FREQUENCY-DOMAIN BLIND SPEECH SEPARATION USING INCOMPLETE DE-MIXING TRANSFORM

Zbyněk Koldovský[1], *Francesco Nesta*[2], *Petr Tichavský*[3], *and Nobutaka Ono*[4]

[1]Faculty of Mechatronics, Informatics, and Interdisciplinary Studies, Technical University of Liberec, Studentská 2, 461 17 Liberec, Czech Republic
[2]Conexant System, 1901 Main Street, Irvine, CA (USA)
[3]Institute of Information Theory and Automation, P.O.Box 18, 182 08 Prague 8, Czech Republic
[4]National Institute of Informatics (NII), SOKENDAI (The Graduate University for Advanced Studies), 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, Japan

ABSTRACT

We propose a novel solution to the blind speech separation problem where the de-mixing transform is estimated only within selected frequency bins. This solution is based on Independent Vector Analysis applied to a subset of instantaneous mixtures, one per selected frequency bin. Next, two approaches are proposed to complete the transform: one based on null beamforming, and the other based on convex programming. In subsequent experiments, we compare combinations of both methods and evaluate their ability to retrieve the whole de-mixing transform. Depending on the number of selected frequencies and the sparsity of room impulse responses, the methods show improvements in terms of computational complexity as well as in terms of separation accuracy.

Index Terms— Blind Source Separation, Independent Vector Analysis, Relative Transfer Function, Sparse Reconstruction, Convex Optimization

1. INTRODUCTION

Spatial acoustic sources are observed on microphones as mixtures of signals that are convolved with acoustic room impulse responses (RIR). Frequency-Domain Blind Source Separation (FDBSS) applies the short-term Fourier transform (STFT) to the mixed signals and treats them as linear instantaneous mixtures, i.e., one mixture per frequency bin [1, 2]. Our goal is to find a linear de-mixing transform that separates a given mixture into the original signals using minimum prior knowledge about the RIRs. The most popular method is Independent Component Analysis, which assumes that the original signals are independent [3].

This work was supported by The Czech Science Foundation through Project No. 14-11898S, by California Community Foundation through Project No. DA-15-114599, and partially by a Grant-in-Aid for Scientific Research (A) (Japan Society for the Promotion of Science (JSPS) KAKENHI Grant Number 16H01735).

This conference paper addresses the basic mixture problem: two sources observed by two microphones, although the ideas could be generalized to more sources and microphones. A 2×2 mixture can be, in the STFT domain, described through

$$\begin{aligned} X_1(k,\ell) &= H_{11}(k)S_1(k,\ell) + H_{12}(k)S_2(k,\ell), \\ X_2(k,\ell) &= H_{21}(k)S_1(k,\ell) + H_{22}(k)S_2(k,\ell), \end{aligned} \tag{1}$$

where $k = 1, \ldots, K$ and $\ell = 1, \ldots, N$ denote, respectively, the frequency and the frame index; K is the length of the Discrete Fourier Transform (DFT); N is the number of STFT frames; X_1 and X_2 denote signals observed on microphones; S_1 and S_2 denote the original signals; and H_{ij} are the transfer functions corresponding to the RIRs. In the vector-matrix form, (1) can be re-written as $\mathbf{X}(k,\ell) = \mathbf{H}(k)\mathbf{S}(k,\ell)$, where $\{\mathbf{H}(k)\}_{ij} = H_{ij}(k)$, $\mathbf{X}(k,\ell) = [X_1(k,\ell), X_2(k,\ell)]^T$, and $\mathbf{S}(k,\ell) = [S_1(k,\ell), S_2(k,\ell)]^T$.

A de-mixing matrix $\mathbf{W}(k)$ is defined such that

$$\mathbf{Y}(k,\ell) = \mathbf{W}(k)\mathbf{X}(k,\ell) = \mathbf{G}(k)\mathbf{S}(k,\ell), \tag{2}$$

where $\mathbf{G}(k) = \mathbf{P}(k)\mathbf{D}(k)$ with $\mathbf{P}(k)$ and $\mathbf{D}(k)$ being, respectively, a permutation and a diagonal matrix. We will refer to the *de-mixing transform* as to the full set of de-mixing matrices $\mathbf{W}(k)$, $k = 1, \ldots, K$, and we will denote it by $\mathbf{W}$.

The uncertainties represented by $\mathbf{P}(k)$ and $\mathbf{D}(k)$ are known in FDBSS, respectively, as the permutation problem and the scaling ambiguity. The permutation problem must be resolved for separating the signals in the time domain [4]. The Minimal Distortion Principle (MDP) is a popular approach to cope with the scaling ambiguity [5]. It reconstructs the separated sources as they are observed on microphones.

FDBSS can be solved by means of Independent Vector Analysis (IVA) where the signals are separated by restoring their independence as with ICA, but the goal here is to also preserve the inter-frequency dependencies. This simultaneously solves the permutation problem [6].

In this paper, we propose to improve the efficiency of IVA by estimating the de-mixing matrices only within selected frequency bins, which yields a de-mixing transform that is *incomplete*. The primary goal is to lower the computational burden with as little loss as possible in separation performance. Speech signals obey sparsity in the frequency domain. It is therefore economical to separate only the active frequencies of the signals.

The secondary goal is to recover the unknown part of the de-mixing transform, which can be useful for the separation of future intervals of the signals, for which the active frequencies can be different. We propose two methods to do so. A simple one is based on the knowledge of time-differences of arrivals (TDOA) of the original signals. The unknown part of the de-mixing transform is completed by null beamformers; see a related work on the combination of beamforming with ICA [7]. The second approach interprets any de-mixing transform as two relative transfer functions (RTFs). Incomplete RTFs are reconstructed through finding their sparsest representations in the time domain, which is motivated by the fact that typical relative impulse responses obey approximate sparsity or compressibility; see [8].

The following section introduces the incomplete solution of the 2×2 FDBSS problem by means of IVA. Section 3 describes the two methods for reconstructing the entire de-mixing transform from its incomplete estimate. Sections 4 and 5 present our experiments and conclude the paper.

2. INCOMPLETE FREQUENCY-DOMAIN BSS

2.1. Equivalence between source separation and source suppression in the 2×2 case

In the 2×2 scenario, the task of separating a source is equivalent to that of suppressing the other source. Therefore, any de-mixing transform can be interpreted through two relative transfer functions (RTF) related to the sources, and vice versa.

Specifically, the RTF related to the ith source is defined as

$$H_{\mathrm{RTF},i}(k) = \frac{H_{2i}(k)}{H_{1i}(k)}. \tag{3}$$

Many methods for estimating the RTF from noise-free or noisy data exist, including those based on BSS; see, e.g., [8, 12, 13, 14]. Let us define $\mathbf{W}$ as

$$\mathbf{W}(k) = \begin{pmatrix} H_{\mathrm{RTF},2}(k) & -1 \\ H_{\mathrm{RTF},1}(k) & -1 \end{pmatrix}. \tag{4}$$

This $\mathbf{W}$ satisfies definition (2) of a de-mixing transform where $\mathbf{G} = \mathtt{diag}(H_{11}H_{\mathrm{RTF},2} - H_{21}, H_{12}H_{\mathrm{RTF},1} - H_{22})$; $\mathtt{diag}(\cdot)$ denotes a diagonal matrix with the argument on its diagonal.

By contrast, any de-mixing matrix $\mathbf{W}(k)$ can be re-scaled, i.e., multiplied by a regular diagonal matrix. Let $\mathbf{\Lambda}(k) = \mathtt{diag}(-1/W_{12}(k), -1/W_{22}(k))$. It can be verified that if $\mathbf{W}(k)$ is de-mixing then $\mathbf{\Lambda}(k)\mathbf{W}(k)$ is equal to (4).

2.2. Independent Vector Analysis

The classical solution of FDBSS applies ICA to each mixture $\mathbf{X}(k,\ell) = \mathbf{H}(k)\mathbf{S}(k,\ell)$ in parallel, by which de-mixing matrices $\mathbf{W}(k)$, $k = 1,\dots,K$, are estimated. Then the permutation problem must be solved to form the whole de-mixing transform; see, e.g., [4, 7, 9, 10].

Independent Vector Analysis solves the estimation and permutation problem simultaneously [6]. It assumes a joint probabilistic model where the frequency components belonging to the same source are mutually dependent. The de-mixing transform is found through minimizing the Kullback-Leibler divergence between the joint probability function and the product of probability densities of vectors where each vector contains the frequency components of the corresponding source.

The natural gradient learning rule [11] for the ijth element of $\mathbf{W}(k)$ reads

$$\mathbf{W}_{ij}(k) \leftarrow \mathbf{W}_{ij}(k) + \mu \sum_{\ell=1}^{2} \Big(\delta_{i\ell} - \mathrm{E}[\phi_k(y_i(1),\dots,y_i(K))\overline{y_\ell(k)}] \Big) \mathbf{W}_{\ell j}(k) \tag{5}$$

where δ denotes the Kronecker delta, $\mathrm{E}[\cdot]$ stands for the expectation operator, μ is a step-length parameter, $y_i(k)$ denotes a random variable modeling the kth frequency component of the ith separated source, and $\phi_k(\cdot)$ is the multivariate score function following from the model for the joint probabilistic density of frequency components of a source. We use the choice from [6] where

$$\phi_k(y_i(1),\dots,y_i(K)) = \frac{y_i(k)}{\sqrt{\sum_{r=1}^{K} |y_i(r)|^2}}. \tag{6}$$

In experiments, we choose $\mu = 0.05$ and perform a fixed number of iterations, namely, 50.

2.3. Estimation of Incomplete De-Mixing Transform

Various natural signals, especially speech, exhibit sparsity in the frequency-domain. The observed signals $\mathbf{X}(k,\ell) = \mathbf{H}(k)\mathbf{S}(k,\ell)$ are dominated by a noise (although the noise term is not explicitly written in the model) when $\mathbf{S}(k,\ell)$ have negligible amplitudes. For such frequencies, the observed data provide little information for estimating the de-mixing matrix.

We propose omitting such frequencies and estimating $\mathbf{W}(k)$ only for $k \in \mathcal{S} \subseteq \{1,\dots,K\}$. A possible solution would be to apply ICA, and perform a permutation correction [4]. Nevertheless, we propose to apply IVA as in [6] but only on a subset of all frequencies. The modification of the above-described algorithm is straightforward: In (5), as well as in (6), k is only allowed to take on the values from $\mathcal{S}$ (as well as the index r in (6)). It follows that the computational

complexity of this modification is $\mathcal{O}(|\mathcal{S}|)$ where $|\mathcal{S}|$ denotes the number of elements in $\mathcal{S}$.

3. DE-MIXING TRANSFORM COMPLETION

For $k \notin \mathcal{S}$, the de-mixing matrices $\mathbf{W}(k)$ could be put equal to zero because of "small" signals' activity within the respective frequency bins. Such a solution would be practical due to a negligible computational burden. However, in an on-line or batch processing regime, the frequencies can be active in future frames. This motivates us to find more sophisticated ways for completing the de-mixing transform.

3.1. Null Beamforming using TDOAs

A simple method proceeds by putting the rows of $\mathbf{W}(k)$, $k \notin \mathcal{S}$, equal to null beamformers where each row steers the spatial null towards a source. Compared to (4), which cannot be used since $H_{\mathrm{RTF},2}(k)$ and $H_{\mathrm{RTF},1}(k)$ are not known, the choice is

$$\mathbf{W}(k) = \begin{pmatrix} e^{-2\pi \mathrm{i} \tau_2 (k-1)/K} & -1 \\ e^{-2\pi \mathrm{i} \tau_1 (k-1)/K} & -1 \end{pmatrix} \tag{7}$$

where τ_1 and τ_2 denote the TDOAs of the first and second source, respectively.

The TDOAs can be estimated from incomplete RTFs [15]. Since the TDOAs are important for efficient initialization of the learning algorithm (5), we assume that their estimates have already been given before applying the IVA and can be used in (7).

3.2. Sparse RTF reconstruction

By being re-scaled, the estimated de-mixing matrices $\mathbf{W}(k)$, $k \in \mathcal{S}$ get the form (4), by which two incomplete RTFs are given. The completion of the de-mixing transform is thus equivalent to the completion of the RTFs. Here we suggest performing the latter through finding the sparsest representations of the incomplete RTFs in the time domain, which was first proposed in [8]. This approach is justified by the fact that the relative impulse response related to the RTF is typically a fast decaying sequence, so it is compressible or approximately sparse.

For a brief description of the method from [8], let $\mathbf{Y}$ be a $|\mathcal{S}| \times 1$ vector collecting the elements of an incomplete RTF H_{RTF}. Its jth element is

$$\mathbf{Y}_j = H_{\mathrm{RTF}}(k_j), \qquad k_j \in \mathcal{S}, \tag{8}$$

where $\mathcal{S} = \{k_1, \ldots, k_{|\mathcal{S}|}\} \subset \{1, \ldots, K\}$.

Finding the sparsest representation of $\mathbf{Y}$ in the time domain is a combinatorial problem, which can be solved through convex relaxation. Here, we use the LASSO formulation [16]

$$\widehat{\mathbf{H}}_{\mathrm{RTF}} = \arg\min_{\mathbf{H}} \|\mathbf{H}_{\mathcal{S}} - \mathbf{Y}\|^2 + \epsilon \|\mathbf{F}^H \mathbf{H}\|_1 \tag{9}$$

where $\epsilon > 0$ controls the sparsity of the solution, $\mathbf{F}$ is the $K \times K$ unitary matrix of the DFT, and the subscript $(\cdot)_{\mathcal{S}}$ denotes a vector/matrix with selected elements/rows whose indices are in $\mathcal{S}$. To find the solution of (9), we use the fast proximal algorithm proposed in [8], whose complexity is $\mathcal{O}(K \log K)$ per iteration[1].

For more advanced formulations of the optimization problem see, e.g., [17, 18].

4. EXPERIMENTS

An experiment with two speech recordings was performed where one recording is a male utterance and the other is a female utterance[2]. Both signals have 10 seconds in length; the sampling frequency is 16 kHz. Signals are processed in the STFT domain with the DFT length of 1024 samples and hop-size 128.

A stereo mixture of these signals was simulated using the room impulse response generator[3]. A room $5 \times 4 \times 3$ m in size was considered, with a reverberation time of $T_{60} = 360$ ms. The speakers were located in the center of the room. Both were one meter distant from two microphones, respectively, at angles of -60° and 60°. The distance between the microphones was 3 cm.

The signals were convolved with the RIRs and mixed together; the initial signal-to-interference ratio (SIR) was ± 0.8 dB. An auxiliary mixture was generated in the same scenario but as if both source signals were white Gaussian sequences; the purpose of this mixture is to evaluate de-mixing transforms with signals that uniformly excite the whole frequency range (the blind estimation will *not* be performed with this mixture).

"Oracle" estimates of the RTFs related to the speakers were computed using the conventional least-squares estimator applied to the responses of the Gaussian sequences. Then, the RTFs were used to construct an "oracle" de-mixing transform through (4).

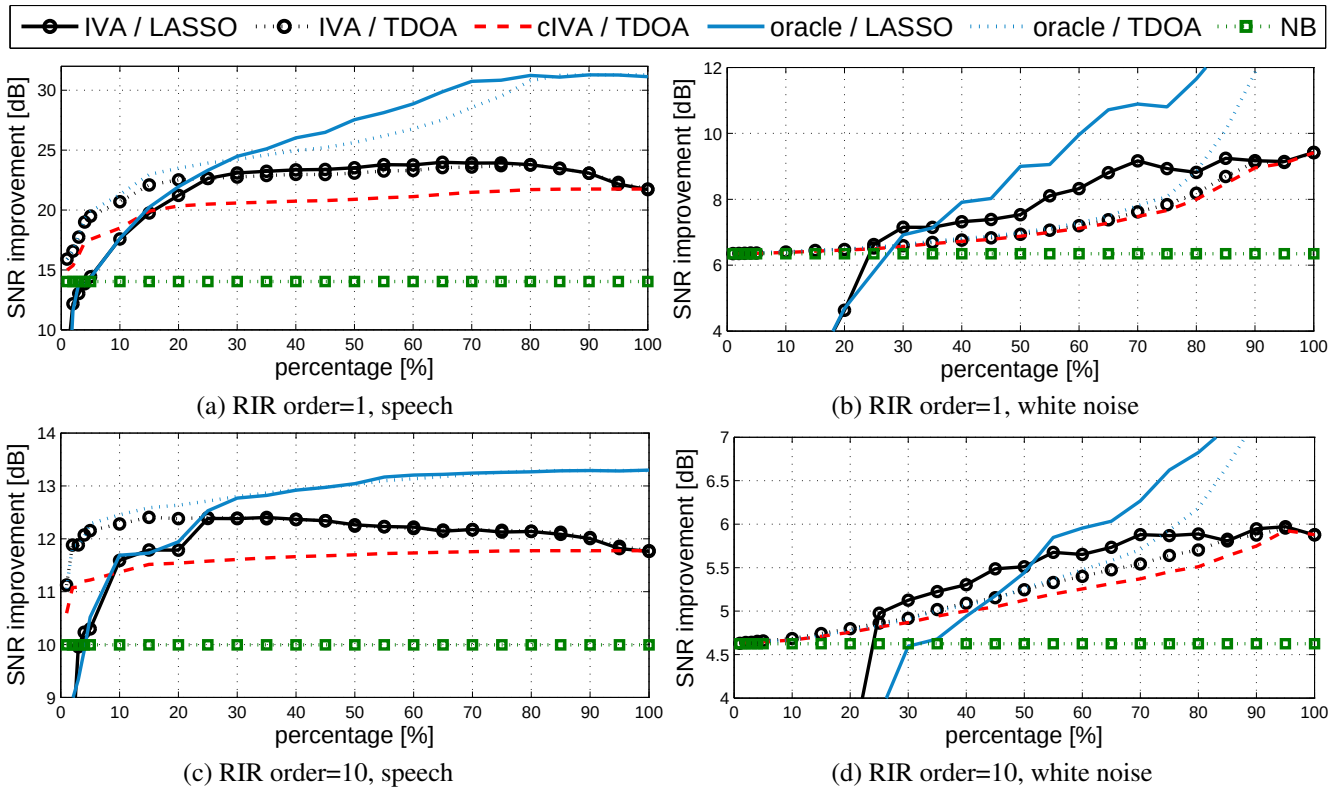
The modified IVA was applied to the mixture of speech signals to estimate the (incomplete) de-mixing transform. First, the percentage p of the most active frequencies within the signal mixture (on average over both channels) was chosen, whereby the set $\mathcal{S}$ was selected; p will be referred to as *percentage*. Second, the modified IVA is applied where (7) is the initialization. Note that, for $p = 100$, the IVA coincides with the original method from [6], here denoted as "cIVA".

When $p < 100$, the resulting de-mixing transform is completed using the two approaches introduced in Section 3. The one based on estimated TDOAs will be denoted as "TDOA".

[1] A Matlab implementation of the algorithm is available at `http://itakura.ite.tul.cz/zbynek/dwnld/SpaRIR.m`.

[2] The recordings are taken from SiSEC 2011, task "Underdetermined-speech and music mixtures," `http://sisec2011.wiki.irisa.fr`.

[3] `https://www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator`



(a) RIR order=1, speech

(b) RIR order=1, white noise

(c) RIR order=10, speech

(d) RIR order=10, white noise

Fig. 1. SIR improvement averaged over both separated sources and channels. The acronym A/B means that method A was used to obtain (incomplete) de-mixing transform, while method B was used to complete it. "NB" denotes the performance of null beamforming (independent of p).

Note that for $p = 0$, the matrix completed by TDOA coincides with null beamforming, which is denoted by "NB". The method based on the sparse reconstruction will be referred to as "LASSO"; in the experiments, (9) is computed with $\epsilon = 0.05$.

After completing the de-mixing transform, the microphone responses of the separated sources are computed and transformed back to the time domain. SIR is evaluated on each channel and the average is taken. Note that this evaluation is, for each de-mixing transform, performed twice: once with the speech mixture, and separately with the Gaussian mixture.

The RIR-generator enables us to select the reflection order of the room. For order = 1, the generated RIRs (and also the respective relative impulse responses) are sparse, so they obey the principal assumption of LASSO. By contrast, when the order is 10, the RIRs contain thousands of reflections and simulate more realistic RIRs.

4.1. Results

Improvements of SIR achieved by several combinations of methods and for the two choices of the reflection-order number are shown in Figures 1(a) through 1(d). The first column corresponds to the evaluation performed on the mixture of speech signals, while the second column corresponds to the white noise evaluation.

Let us mention some important facts: For $p = 100$, IVA coincides with cIVA. The methods to complete the de-mixing transform are not applied in this case, so, e.g., IVA/LASSO and IVA/TDOA coincide, etc. The best performance is provided by the oracle method, then IVA and NB. For p below a critical value between 20-30% and, especially, for p close to 0, LASSO fails while TDOA approaches NB. We shall not discuss results for p below the critical value in the rest of this text.

Now, the discussion must be performed separately for the results evaluated on the speech and white noise signals. We begin with the former results (Figures 1(a) and 1(c)).

Here, oracle/LASSO achieves better SIR than oracle/TDOA when p is between 30 and 80 (Fig. 1(a)), but only when the reflection order is 1. The differences between IVA/LASSO and IVA/TDOA are negligible. This follows from the fact that the de-mixing transform is already well identified within the active frequencies of speech, so the SIR cannot be much improved through completing the de-mixing transform.

A positive observation is that an incomplete IVA is slightly improving with decreasing p (until $p \approx 30$). Both

IVA/LASSO and IVA/TDOA outperform cIVA/TDOA for $p < 100$. cIVA thus appears to be inefficient compared to incomplete IVA when p is sufficiently high, both in terms of speed and accuracy.

The SIR achieved with the white Gaussian signals (Figures 1(b) and 1(d)) evaluate the whole de-mixing transform (after completion) uniformly on the whole frequency range. Here, the SIR values achieved by oracle/LASSO and IVA/LASSO when p is between 30% and 90% are significantly better compared to those obtained when TDOA is used instead of LASSO. This phenomenon is more distinctly seen in Fig. 1(b) where the RIRs are more sparse. The differences are lower in Fig. 1(d) where the reflection order is 10. In the latter case, IVA outperforms cIVA as well.

5. CONCLUSIONS

We have shown advantageous properties of FDBSS when performed via incomplete IVA using the natural gradient algorithm. Namely, the computational burden is lower as it grows linearly with p, while the achieved SIR could be even slightly higher provided that p is higher than the critical value. This value could be chosen based on the number of active frequency bins. Results of our experiments are indicative of the fact that LASSO is able to gain some information about demixing matrices outside the active frequencies, depending on the sparsity of the RIRs.

A comprehensive comparison of the proposed concept in combination with other BSS methods will be a subject of our future work.

6. REFERENCES

[1] S. Makino, Te-Won Lee, and H. Sawada, *Blind Speech Separation*, Springer, Sept. 2007.

[2] M. S. Pedersen, J. Larsen, U. Kjems, L. C. Parra, "A survey of convolutive blind source separation methods," in *Springer Handbook of Speech*, Nov. 2007.

[3] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, Wiley-Interscience, New York, 2001.

[4] H. Sawada, R. Mukai, S. Araki, and S. Makino, "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE Trans. Speech Audio Processing*, vol. 12, no. 5, pp. 530-538, 2004.

[5] K. Matsuoka and S. Nakashima, "Minimal distortion principle for blind source separation," *Proceedings of 3rd International Conference on Independent Component Analysis and Blind Source Separation (ICA '01)*, pp. 722-727, San Diego, Calif, USA, Dec. 2001.

[6] T. Kim, H. T. Attias; T.-W. Lee, "Blind Source Separation Exploiting Higher-Order Frequency Dependencies," *IEEE Trans. ASLP*, vol. 15, no. 1, Jan. 2007.

[7] H. Saruwatari, T. Kawamura, T. Nishikawa, A. Lee, and K. Shikano, "Blind Source Separation Based on a Fast-Convergence Algorithm Combining ICA and Beamforming," *IEEE Trans. ASLP*, vol. 14, no. 2, Mar. 2006.

[8] Z. Koldovský, J. Málek, and S. Gannot, "Spatial Source Subtraction Based on Incomplete Measurements of Relative Transfer Function," *IEEE/ACM Trans. on Speech, Audio and Language Processing*, vol. 23, no. 8, pp. 1335–1347, Aug. 2015.

[9] F. Nesta and M. Matassoni, "Blind source extraction for robust speech recognition in multisource noisy environments," *Comput. Speech Lang.*, vol. 27, no. 3, pp. 703-725, May 2013.

[10] P. Sudhakar and R. Gribonval, "A sparsity-based method to solve the permutation indeterminacy in frequency domain convolutive blind source separation," *ICA 2009*, pp. 338–345, Paraty, Brazil, March 2009.

[11] S.-I. Amari, A. Cichocki, and H. H. Yang, "A New Learning Algorithm for Blind Signal Separation," *Adv. Neural Inf. Process. Syst.*, vol. 8, pp. 752–763, 1996.

[12] O. Shalvi and E. Weinstein, "System Identification Using Nonstationary Signals," *IEEE Trans. Signal Processing*, vol. 44, no. 8, pp. 2055–2063, Aug. 1996.

[13] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Trans. on Signal Processing*, vol. 49, no. 8, pp. 1614–1626, Aug. 2001.

[14] I. Cohen, "Relative Transfer Function Identification Using Speech Signals," *IEEE Trans. ASLP*, vol. 12, no. 5, Sept. 2004.

[15] D. Malioutov, M. Cetin, and A. S. Willsky, "A sparse signal reconstruction perspective for source localization with sensor arrays," *IEEE Trans. Signal Process.*, vol. 53, no. 8, pp. 3010-3022, Aug. 2005.

[16] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Royal. Statist. Soc. B.,* vol. 58, pp. 267–288, 1996.

[17] Z. Koldovský and P. Tichavský, "Sparse Reconstruction of Incomplete Relative Transfer Function: Discrete and Continuous Time Domain," *EUSIPCO 2015*, pp. 394–398, Nice, France, Sept. 2015.

[18] Z. Koldovský, J. Málek, and P. Tichavský, "Improving Relative Transfer Function Estimates Using Second-Order Cone Programming," *LVA/ICA 2015*, pp. 227–234, Liberec, Czech Republic, Aug. 2015.