A Local Model of Relative Transfer Functions Involving Sparsity

Zbyněk Koldovský[1], Jakub Janský[1], and Francesco Nesta[2]

[1] Faculty of Mechatronics, Informatics and Interdisciplinary Studies
Technical University of Liberec, Studentská 2, 461 17 Liberec, Czech Republic
jakub.jansky@tul.cz, zbynek.koldovsky@tul.cz,
[2] Conexant System, 1901 Main Street, Irvine, CA (USA),
Francesco.Nesta@conexant.com

Abstract. We propose a model of Relative Transfer Functions between two microphones which correspond to closed target positions within a certain spatially constrained area. Each RTF is modeled as the product of two transfer functions. One corresponds to a linear-phase filter and is the common factor of all the RTFs. The second transfer function is an individual factor that should be as sparse as possible in the time domain. A learning algorithm to identify the decomposition given a set of RTFs is proposed. The common factor is the main output, which we then apply to reconstruct an unknown RTF corresponding to a position within the assumed area, when only an incomplete measurement of it is available.

Keywords: Relative Transfer Function, Generalized Sidelobe Canceler, Convex Programming, Sparse Modeling

1 Introduction

A noisy stereo recording of a target signal can be, in the short-term Discrete Fourier Transform (DFT) domain, described as

$$\begin{aligned} X_{\mathrm{L}}(k,\ell) &= S_{\mathrm{L}}(k,\ell) + Y_{\mathrm{L}}(k,\ell), \\ X_{\mathrm{R}}(k,\ell) &= H(k)S_{\mathrm{L}}(k,\ell) + Y_{\mathrm{R}}(k,\ell), \end{aligned} \tag{1}$$

where k and ℓ denote, respectively, the frequency and the frame index; let the DFT window length be M. Further let X_{L} and X_{R} denote signals observed by microphones; S_{L} is the response (image) of the target signal on the left microphone; Y_{L} and Y_{R} are the remaining signals (noise and interferences) commonly referred to as noise. H is the Relative Transfer Function (RTF) between the microphones related to the target signal.

The RTF is an important component of multichannel audio signal processing systems [1]. Using the RTF, a multichannel filter can be designed such that it

[0] This work was supported by California Community Foundation through Project No. DA-15-114599.

performs spatial null towards a target source, thereby yielding a noise signal reference on its output [3]. Specifically, let $\widehat{H}$ be an estimate of H and the multichannel filter with inputs X_{L} and X_{R} be such that its output is $Z(k,\ell) = \widehat{H}(k)X_{\mathrm{L}}(k,\ell) - X_{\mathrm{R}}(k,\ell)$. By (1), it holds that

$$Z(k,\ell) = \underbrace{\big(\widehat{H}(k) - H(k)\big)S_{\mathrm{L}}(k,\ell)}_{\text{target signal leakage}} + \underbrace{\widehat{H}(k)Y_{\mathrm{L}}(k,\ell) - Y_{\mathrm{R}}(k,\ell)}_{\text{noise reference}}. \tag{2}$$

For $\widehat{H} = H$, the target signal leakage vanishes, and $Z(k,\ell) = H(k)Y_{\mathrm{L}}(k,\ell) - Y_{\mathrm{R}}(k,\ell)$ provides the key noise reference signal.

However, real-world RTFs have many coefficients that can change quickly in a short period of time. The RTF estimation thus poses a difficult problem. Several estimators have been proposed to estimate the RTF directly from noisy recordings. A frequency-domain estimator assuming nonstationary target signal and stationary noise was proposed in [4]. Methods based on Blind Source Separation (BSS) and Independent Component Analysis (ICA) can cope with general situations (e.g., with nonstationary directional interfering sources) [5]. However, the accuracy of such estimation is limited.

During intervals where only the target source is active, conventional least-square approaches can be used to obtain highly accurate RTF estimates. These estimates can be used later when noise is active. However, the acoustic conditions must remain the same: in particular, the position of the target source must be preserved. The fact that the target source position is often limited to a confined area can be used as a priori knowledge. For example, it is possible to collect a bank of RTFs during noise-free intervals such that the area is covered by the bank [2]. The problem of estimating the RTF can then be simplified to one of choosing an appropriate RTF from the bank.

Recent methods aim to find suitable low-rank models for such banks of acoustic transfer functions, which are instrumental in computing highly accurate RTF estimates from noisy recordings; see, e.g., [6, 7]. In this paper, we propose a novel sparsity-based model, which is applied when reconstructing incomplete RTF (iRTF), that is, an RTF estimate whose values are known only for certain frequencies. In [8], the iRTF is completed through finding its sparsest representation in the time domain. This is justified by the fact that relative impulse responses (ReIRs), i.e., the time-domain counterparts of RTFs, are compressible (approximately sparse) sequences. However, since exact sparsity is invoked in [8], there are performance limitations; see also [9]. The goal of our paper is to exploit the proposed model in order to lower this performance loss.

Throughout the paper, upper-case letters will denote transfer functions (DFT domain) while their time-domain counterparts will be denoted by lower-case letters. Bold letters will denote vectors comprising coefficients of the corresponding quantities. For example, $(\mathbf{H})_k = H(k)$. The time domain counterpart of H, called the relative impulse response (ReIR), is h, and $(\mathbf{h})_k = h(k)$.

2 Model Proposal

Let H_p be the RTF for the pth position of the target source within a confined area, and $p = 1, \dots, P$; h_p be the corresponding ReIR. The positions $1, \dots, P$ can form either a regular or an irregular grid [2, 12]. Our goal is to model H_p as

$$H_p(k) \approx B(k)G_p(k), \quad k = 0, \dots, M-1, \quad p = 1, \dots, P, \tag{3}$$

where g_p is sparse. B is independent of the position index p, so it is a common factor of $H_1, \dots, H_P$.

As a motivator, note that each h_p can be approximated by a sparse g_p such that neither $\|\mathbf{h}_p - \mathbf{g}_p\|_2$ nor the energy of the target signal leakage in (2) for $\widehat{H} = G_p$ is higher than a chosen limit; see [10]. The conjecture behind the proposed model is that the residual $B = H_p/G_p$ might be independent of p since the RTFs come from the same area. Assuming that B is known, an iRTF for a target position within the area can be reconstructed as $B \cdot G$ by invoking the sparsity of g.

Linear-phase unit-norm constraint Definition (3) is not unique. The scale of B can be arbitrarily changed, which can be compensated for by the reciprocal scaling of all G_p, $p = 1, \dots, P$. To solve the scaling uncertainty, B will be constrained to have a unit norm. Similarly, B can be arbitrarily delayed. Therefore, we constrain B to have a constant group delay of $\lfloor M/2 \rfloor$, hence, linear phase. It means that b is constrained to be symmetric[3] along its $(\lfloor M/2 \rfloor + 1)$th coefficient; we will assume, without loss on generality, that M is odd. The set of the unit-norm symmetric filters (vectors) of length M will be denoted by $\mathcal{A}_M$.

A method to learn the common factor B Let $H_1, \dots, H_P$ be given. Before learning the common factor B, the causality of the filters in (3) must be ensured by delaying each h_p by $\lfloor M/2 \rfloor + D$ samples. The delay by $\lfloor M/2 \rfloor$ is due to the linear-phase constraint imposed on B. The value of $D \geq 0$ must be sufficiently high to ensure the causality of g_ps; we typically use $D = 10$.

To find B, we formulate an optimization problem with the above constraints on B as

$$\mathbf{B} = \arg \min_{\mathbf{B}, \mathbf{G}_1, \dots, \mathbf{G}_P} \sum_{p=1}^{P} \|\mathbf{g}_p\|_0 \quad \text{s.t.} \quad \sum_{p=1}^{P} \|\mathbf{H}_p - \mathtt{diag}(\mathbf{B})\mathbf{G}_p\|_2^2 \leq \epsilon, \quad \mathbf{b} \in \mathcal{A}_M. \tag{4}$$

where $\mathtt{diag}(\cdot)$ denotes a diagonal matrix with the argument on its main diagonal; $\|\cdot\|_0$ denotes the ℓ_0 pseudo-norm that is equal to the number of nonzero elements

[3] This constraint excludes the set of anti-symmetric linear-phase filters. However, b tends to be close to the pure delay filter in practice, hence it is always symmetric. In experiments with real-world RTFs, we did not observe any cases where b was anti-symmetric.

of the argument; ϵ is a free positive constant. Note that $\mathbf{B} = \mathbf{F}\mathbf{b}$, $\mathbf{H}_p = \mathbf{F}\mathbf{h}_p$, and $\mathbf{G}_p = \mathbf{F}\mathbf{g}_p$ where $\mathbf{F}$ is the $M \times M$ unitary matrix of DFT.

However, the problem (4) is NP hard and *not* convex in general even if the ℓ_0 pseudo-norm is replaced by the ℓ_1 norm. Without a guarantee to find the global minimum of (4), we propose an alternating algorithm that can be used to find a satisfactory solution.

In the beginning, it is assumed that feasible $B, G_1, \ldots, G_P$ are given as an initial guess. This can be, for example, $B(k) = e^{\frac{i2\pi\lfloor M/2\rfloor k}{M}}$ and $G_p(k) = \overline{B(k)}H_p(k)$, $\forall p, k$, where $\bar{\cdot}$ denotes the complex conjugate. Then, one iteration of the algorithm consists of two optimization steps. First, all $\mathbf{G}_p$ are fixed while $\mathbf{B}$ is computed such that it minimizes the constraint in (4); that is, the solution of

$$\min_{\mathbf{B}} \sum_{p=1}^{P} \|\mathbf{H}_p - \mathtt{diag}(\mathbf{B})\mathbf{G}_p\|_2^2 \quad \text{s.t.} \quad \mathbf{b} \in \mathcal{A}_M. \tag{5}$$

This is a constrained least-squares problem that can be solved analytically. Let $\mathbf{c} = [c_1, \ldots, c_{\lfloor M/2\rfloor}]^T$ denote the symmetric part of $\mathbf{b}$, i.e., $\mathbf{b} = [\mathbf{c}; b_{\lfloor M/2\rfloor+1}; \underline{\mathbf{c}}]$ where $\underline{\cdot}$ denotes the upside down operator. For $\mathbf{b}$ being the solution of (5) it holds that

$$[\mathbf{c}; b_{\lfloor M/2\rfloor+1}] \propto \left(\mathbf{A}^H \mathtt{diag}\Big(\sum_{p=1}^{P} \mathtt{diag}(\mathbf{G}_p)\overline{\mathbf{G}_p}\Big)\mathbf{A}\right)^{-1} \mathbf{A}^H \left(\sum_{p=1}^{P} \mathtt{diag}(\overline{\mathbf{G}}_p)\mathbf{H}_p\right), \tag{6}$$

where $\mathbf{A}$ is an $M \times (\lfloor M/2\rfloor + 1)$ matrix whose ith row is $[\mathbf{F}_{i,1} + \mathbf{F}_{i,M}, \mathbf{F}_{i,2} + \mathbf{F}_{i,M-1}, \ldots, \mathbf{F}_{i,\lfloor M/2\rfloor} + \mathbf{F}_{i,\lfloor M/2\rfloor+2}, \mathbf{F}_{i,\lfloor M/2\rfloor+1}]$. After computing $\mathbf{b}$ using (6), the vector is normalized to satisfy the constraint $\|\mathbf{b}\|_2 = 1$, so finally $\mathbf{b} \in \mathcal{A}_M$. Note that $B, G_1, \ldots, G_P$ remain feasible for (4) after the first step.

The goal of the second step is to improve the sparsity of $G_1, \ldots, G_P$ while preserving their feasibility. B is fixed while $G_1, \ldots, G_P$ are optimized to approach the solution of (4). To this end, P independent convex programs are solved, one for each G_p,

$$\min_{\mathbf{G_p}} \|\mathbf{g}_p\|_1 \quad \text{w.r.t.} \quad \|\mathbf{H}_p - \mathtt{diag}(\mathbf{B})\mathbf{G}_p\|_2^2 \le \epsilon/P. \tag{7}$$

This optimization problem is well-known under the name of *basis pursuit denoising* (BPDN) and can be efficiently solved using, e.g., SPGL1[4]; see [11]. Using the fact that $\mathbf{F}$ is a unitary matrix, (7) can be written in its equivalent form

$$\min_{\mathbf{g_p}} \|\mathbf{g}_p\|_1 \quad \text{w.r.t.} \quad \|\mathbf{h}_p - \mathbf{F}^H \mathtt{diag}(\mathbf{B})\mathbf{F} \cdot \mathbf{g}_p\|_2^2 \le \epsilon/P. \tag{8}$$

Since $\mathbf{b}$ is real-valued, the matrix $\mathbf{F}^H \mathtt{diag}(\mathbf{B})\mathbf{F}$ as well as the whole program (8) are real-valued. The proposed optimization algorithm is summarized in Algorithm 1.

[4] `http://www.cs.ubc.ca/~mpf/spgl1`

Algorithm 1: Learning algorithm to find B.

Input: $\mathbf{H}_1, \ldots, \mathbf{H}_P$ and ϵ
Output: $\mathbf{B}$
Initialization: $(\mathbf{B})_k = e^{\frac{i2\pi \lfloor M/2 \rfloor k}{M}}$, $(\mathbf{G}_p)_k = \overline{(\mathbf{B})_k}(\mathbf{H}_p)_k$, $\mathbf{B}_{\text{old}} = \mathbf{0}_{M\times 1}$
while $\|\mathbf{B}_{old} - \mathbf{B}\|_2 > \texttt{tol}$ **do**
  $\mathbf{B}_{\text{old}} = \mathbf{B}$;
  Compute $\mathbf{b}$ using (6);
  $\mathbf{b} \leftarrow \mathbf{b}/\|\mathbf{b}\|_2$;
  $\mathbf{B} = \texttt{fft}(\mathbf{b})$;
  $\mathbf{Q} = \mathbf{F}^H \texttt{diag}(\mathbf{B})\mathbf{F}$;
  for $p \in \{1, \ldots, P\}$ **do**
    $\mathbf{g}_p = \arg\min_{\mathbf{g}} \|\mathbf{g}\|_1 \quad \text{w.r.t.} \quad \|\mathbf{h}_p - \mathbf{Q}\,\mathbf{g}\|_2 \leq \epsilon/P$;
    $\mathbf{G}_p = \texttt{fft}(\mathbf{g}_p)$;
  end
end

3 Application: Sparse Recovery of an Incomplete RTF

An iRTF is represented by an $|\mathcal{S}| \times 1$ vector $\mathbf{Y}$ whose kth element is

$$(\mathbf{Y})_k = \widehat{H}(i_k), \qquad k \in \mathcal{S}, \tag{9}$$

where $\mathcal{S} = \{i_1, \ldots, i_{|\mathcal{S}|}\} \subset \{1, \ldots, M\}$ is the set of indices of known values of $\widehat{H}$. In [8], it is proposed to retrieve the complete RTF estimate from $\mathbf{Y}$ through

$$\widehat{\mathbf{h}} = \arg\min_{\mathbf{h}} \|\mathbf{h}\|_1 \quad \text{w.r.t.} \quad \|\mathbf{Y} - \mathbf{F}_{\mathcal{S}}\mathbf{h}\|_2 \leq \epsilon, \tag{10}$$

where the subscript $(\cdot)_{\mathcal{S}}$ denotes a vector/matrix with selected rows whose indices are in $\mathcal{S}$; ϵ is a free positive parameter[5]. This is the BPDN optimization program, which, in other words, seeks for the sparsest representation of $\mathbf{y}$ in the time domain. As mentioned in the Introduction, the performance of this method is limited due to the fact that the original h is typically not exactly sparse.

Now, assume that $\widehat{H}$ is a noisy estimate of RTF for the current (unknown) position of the target within an assumed area. Let $\mathbf{Y}$ be its incomplete version where $\mathcal{S}$ is selected on the basis of a certain hypothesis (e.g., we select only those elements of $\widehat{H}$ that are not affected by the noise [8]). Next, assume that a bank of RTFs $H_1, \ldots, H_P$ is given. These RTFs are valid for some positions within the area but can be different from the current ones estimated by $\widehat{H}$. Using the bank of RTFs, B from (3) can be identified using Algorithm 1 with some ϵ. The goal is now to exploit B as a priori knowledge when retrieving H from $\mathbf{Y}$.

We propose computing the complete RTF estimate as

$$\widehat{\mathbf{H}} = \texttt{diag}(\mathbf{B})\mathbf{G} \tag{11}$$

[5] In fact, $\widehat{\mathbf{h}}$ is in [8] sought through solving $\min_{\mathbf{h}} \|\mathbf{F}_{\mathcal{S}}\mathbf{h} - \mathbf{Y}\|_2^2 + \tau\|\mathbf{h}\|_1$ where $\tau > 0$. This is, nevertheless, an equivalent problem to (10) in the sense that there exists τ such that the solutions are the same.

where $\mathbf{G} = \mathbf{F}\mathbf{g}$, and

$$\mathbf{g} = \arg\min_{\mathbf{g}} \|\mathbf{g}\|_1 \quad \text{w.r.t.} \quad \|\mathbf{Y} - \mathtt{diag}(\mathbf{B}_{\mathcal{S}})(\mathbf{F}_{\mathcal{S}}\mathbf{g})\| \leq \epsilon. \tag{12}$$

Discussion In view of the Compressed Sensing theory, (10) as well as (12) can be interpreted as sparse reconstructions of compressed measurements $\mathbf{Y}$ in the time domain when the sensing matrix is, respectively, $\mathbf{F}_{\mathcal{S}}$ and $\mathtt{diag}(\mathbf{B}_{\mathcal{S}})\mathbf{F}_{\mathcal{S}}$. In other words, the sensing domain is, respectively, the Fourier domain and the Fourier domain transformed through $\mathtt{diag}(\mathbf{B})$. The principal difference depends on the distance of $\mathbf{B}$ from the pure delay by $\lfloor M/2 \rfloor$ samples.

4 Experimental Verification

The first experiment was based on simulated data. An artificial RTF of length $M = 1025$ was generated as $H = B \cdot G$ where $(\mathbf{g})_k = a_k e^{-0.008(|k|-20)}$, $k = 1, \ldots, M$, with only q random nonzero a_k from $[-0.5; 0.5]$, and $a_{20} = 1$. Hence, $\mathbf{g}$ is q-sparse and has an exponential decay with the highest peak at $k = 20$. B was generated as $(\mathbf{B})_k = d_k e^{\frac{i2\pi k \lfloor M/2 \rfloor}{M}}$ where d_k were taken at random from $[0.5; 1.5]$. Such B has a linear phase and is close to the pure delay by $\lfloor M/2 \rfloor$ samples.

A 10 s female voice signal was taken to simulate a noise-free recording according to (1). Then, H was estimated from each 1 s interval of the recording using the least-squares estimator. Then only p percent of the most active frequencies in the original signal were put in $\mathcal{S}$, and the iRTF was created. This procedure simulates situations when the signal by the target source does not excite the full frequency range or when some frequency bins are contaminated by noise. Therefore, only iRTF is available. The reconstructed estimates of H were computed through (10) and (12) using known B, respectively, both with $\epsilon = 0.1$; the sampling frequency was 16 kHz.

The resulting RTFs were evaluated in terms of attenuation rate (ATR). The ATR is defined as the ratio between the power of the original signal on the left microphone and the power of the target signal leakage term in (2). The more negative the ATR (in dB), the better the target signal blocking. The resulting ATRs averaged over the respective intervals are shown in Fig. 1. The "model-based solution" through (12) using known B achieves significantly better ATR than the sparse solution by (10) for all levels of the incompleteness p. For $p \to 100$, both solutions approach the ATR by the true RTF.

The second experiment was conducted with real-world recordings. In an office room with $T_{60} \approx 300$ ms, a 12 s sequence of audio signals (4 s white noise + 4 s male speech + 4 s female speech) played by a loudspeaker was recorded by two microphones. The loudspeaker was placed in front of the microphones at a distance of 1.5 m and rotated at nine different angles from -90 through $90°$ ($0°$is the direction towards the microphones). The microphone spacing was 5 cm.

The RTFs for the loudspeaker's rotations were computed using the first second of the recorded white noise and the least-square estimator (from this point forward referred to as "true RTF"). Algorithm 1 was applied with $\epsilon = 0.01$ and

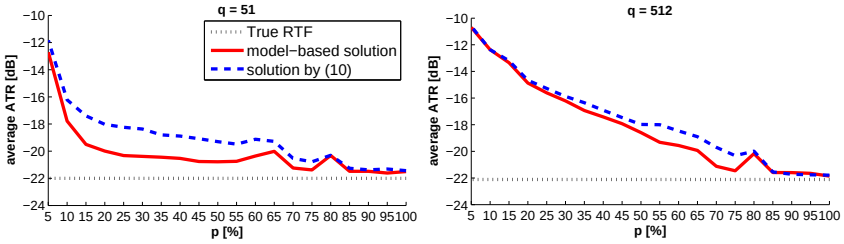

Fig. 1. Results of the first experiment for two levels of sparsity of the factor G in $H = B \cdot G$, namely, $q = 51$ and $q = 512$.

$\texttt{tol} = 10^{-3}$ to learn the common factor B of true RTFs corresponding to the rotations by $-30°$, $0°$, and $30°$. The method converged after 46 iterations.

The recordings were divided into twelve 1 s intervals. On each interval, the least-square RTF estimate was computed and 10 percent of its values corresponding to the most active frequencies on the left microphone was selected to build up the iRTF.

The relative ATRs averaged over 4-s intervals are shown in Fig. 2. These results show that the knowledge of B helps to reconstruct the RTF with the average relative improvement of ATR by up to 2 dB. It should be noted that although B was derived only using known RTFs for rotations by $-30°$, $0°$, and $30°$, no overlearning is observed in Fig. 2 for these rotations: The ATR improvement is rather uniform over all rotations and mainly depends on the original signal. Unlike white noise, the speech signals are easier to attenuate by the reconstructed RTF as they do not span the whole frequency range.

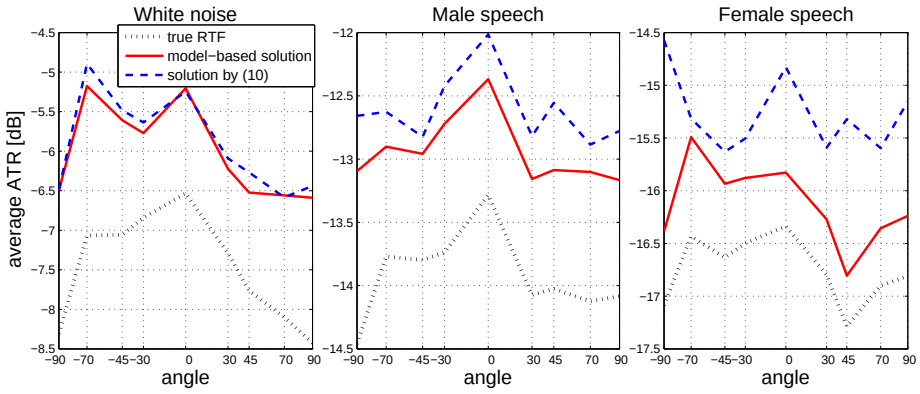

Fig. 2. Average ATR achieved by RTFs reconstructed using (10) and the model-based solution (12). The ATR achieved by the true RTF is shown and corresponds to an optimum achievable performance.

5 Conclusions

Algorithm 1 was shown to be able to learn the common factor of RTFs that are known for positions of a target source within a confined area. This factor can be used when reconstructing an incomplete measurement of an RTF from the same area. An interpretation in terms of the Compressed Sensing theory is that the proposed method learns a new sensing domain (or, equivalently, a sparsity domain) of the RTFs through finding their common factor.

References

1. Benesty, J., Makino, S., and Chen, J. (Eds.), *Speech Enhancement*, 1st edition, Springer-Verlag, Heidelberg, 2005.
2. Koldovský, Z., Málek, J., Tichavský, P., and Nesta, F., "Semi-blind Noise Extraction Using Partially Known Position of the Target Source," IEEE Trans. on Speech, Audio and Language Processing, 21(10):2029–2041, Oct. 2013.
3. Gannot, S., Burshtein, D., and Weinstein, E., "Signal Enhancement Using Beamforming and Nonstationarity with Applications to Speech," IEEE Trans. on Signal Processing, 49(8):1614–1626, Aug. 2001.
4. Shalvi, O., Weinstein, E., "System Identification Using Nonstationary Signals," IEEE Trans. Signal Processing, 44(8):2055–2063, Aug. 1996.
5. Reindl, K., Markovich-Golan, S., Barfuss, H., Gannot, S., Kellermann, W., "Geometrically Constrained TRINICON-based relative transfer function estimation in underdetermined scenarios," IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), pp. 1–4, 2013.
6. Talmon, R., Gannot, S., "Relative Transfer Function Identification on Manifolds for Supervised GSC Beamformers," Proc. of the 21st European Signal Processing Conference (EUSIPCO), Marrakech, Morocco, Sep. 2013.
7. Haneda, Y., Makino, S., Kaneda, Y., "Common Acoustical Pole and Zero Modeling of Room Transfer Functions," IEEE Trans. on Speech and Audio Processing, 2(2):320–328, April 1994.
8. Koldovský, Z., Málek, J., and Gannot, S., "Spatial Source Subtraction Based on Incomplete Measurements of Relative Transfer Function," accepted in IEEE Trans. on Speech, Audio and Language Processing, April 2015.
9. Koldovský, Z., Tichavský, P., "Sparse Reconstruction of Incomplete Relative Transfer Function: Discrete and Continuous Time Domain," accepted for a special session at EUSIPCO 2015, Nice, France, Sept. 2015.
10. Málek, J., Koldovský, Z., "Sparse Target Cancellation Filters with Application to Semi-Blind Noise Extraction," Proc. of the 41st IEEE International Conference on Audio, Speech, and Signal Processing (ICASSP 2014), Florence, Italy, pp. 2109–2113, May 2014.
11. Berg, E. van den, Friedlander, M. P., "Probing the Pareto frontier for basis pursuit solutions," SIAM J. on Scientific Computing, 31(2):890–912, Nov. 2008.
12. Málek, J., Botka, D., Koldovský, Z., and Gannot, S., "Methods to Learn Bank of Filters Steering Nulls toward Potential Positions of a Target Source," Proc. of the 4th Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA 2014), Nancy, France, May 12–14, 2014.