Improving Relative Transfer Function Estimates Using Second-Order Cone Programming

Zbyněk Koldovský[1,2], Jiří Málek[1], and Petr Tichavský[2]

[1] Faculty of Mechatronics, Informatics and Interdisciplinary Studies
Technical University of Liberec, Studentská 2, 461 17 Liberec, Czech Republic
zbynek.koldovsky@tul.cz,
[2] Institute of Information Theory and Automation, Pod vodárenskou věží 4,
P.O. Box 18, 182 08 Praha 8, Czech Republic

Abstract. This paper addresses the estimation of Relative Transfer Function (RTF) between microphones from noisy recordings. We utilize an incomplete initial measurement of the RTF, which is known for only several frequency bins. The measurement is completed by finding its sparsest representation in the time domain. We propose to perform this reconstruction by solving a Second-Order Cone Program (SOCP). Free parameters of this formulation represent distance of the completed RTF from the initial estimate. We select these parameters based on the theoretical performance of the initial estimate. In experiments with real-world data, this approach achieves a significant refinement of the RTF, especially in scenarios with low signal-to-noise ratios.

Keywords: Audio Signal Processing, Relative Transfer Function, Compressed Sensing, Second-Order Cone Programming, Incomplete Measurement

1 Introduction

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

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

where k and ℓ denote, respectively, the frequency and the frame index; let the DFT length be M; S denotes the target signal; X_{L} and X_{R} correspond, respectively, to the signals observed on the left and right microphones; Y_{L} and Y_{R} are the remaining signals (noise and interferences) commonly referred to as noise. H_{L} and H_{R} denote the acoustic transfer functions between the microphones and

[2] This work was supported by The Czech Sciences Foundation through Project No. 14-11898S.

the target, which are assumed to be approximately constant (independent of ℓ) during short intervals.

Define the relative transfer function (RTF) as $H_{\mathrm{RTF}}(k) = H_{\mathrm{R}}(k)H_{\mathrm{L}}(k)^{-1}$. Then, (1) can be re-written as

$$\begin{aligned} X_{\mathrm{L}}(k,\ell) &= S_{\mathrm{L}}(k,\ell) + Y_{\mathrm{L}}(k,\ell) \\ X_{\mathrm{R}}(k,\ell) &= H_{\mathrm{RTF}}(k)S_{\mathrm{L}}(k,\ell) + Y_{\mathrm{R}}(k,\ell) \end{aligned} \tag{2}$$

where $S_{\mathrm{L}}(k,\ell) = H_{\mathrm{L}}(k)S(k,\ell)$. The time domain counterpart of H_{RTF}, called the relative impulse response (ReIR), will be denoted as h_{rel}.

Knowing H_{RTF} (or h_{rel}) enables to design an efficient spatial filter with two inputs X_{L} and X_{R} such that it cancels the target source and only pass through the noise signals. The output of the spatial filter[3] is $Z(k,\ell) = H(k)X_{\mathrm{L}}(k,\ell) - X_{\mathrm{R}}(k,\ell)$, which is determined by the transfer function H. By (2), it holds that

$$Z(k,\ell) = \underbrace{\big(H(k) - H_{\mathrm{RTF}}(k)\big)S_{\mathrm{L}}(k,\ell)}_{\text{target signal leakage}} + \underbrace{H(k)Y_{\mathrm{L}}(k,\ell) - Y_{\mathrm{R}}(k,\ell)}_{\text{noise reference}}. \tag{3}$$

For $H = H_{\mathrm{RTF}}$ the target signal leakage vanishes, and $Z(k,\ell) = H_{\mathrm{RTF}}(k)Y_{\mathrm{L}}(k,\ell) - Y_{\mathrm{R}}(k,\ell)$. Hence, $Z(k,\ell)$ provides the key noise reference signal, which is important for audio applications such as source separation or speech enhancement.

The signal-to-noise ratio (SNR) in $Z(k,\ell)$ can be used as a practical evaluator of $H(k)$. We will therefore use *attenuation ratio* (ATR), which is the ratio between the initial SNR in (1) and the SNR in $Z(k,\ell)$.

The RTF estimation when noise is active is a challenging problem. During noise-free intervals, conventional time-domain or frequency-domain estimators can be used. The obtained RTF can be used later when noise is active, however, the position of the target must remain the same. To estimate the RTF from noisy data, Shalvi and Weinstein proposed a method assuming model where nonstationary target signal is interfered by a stationary noise [2]. Methods based on Blind Source Separation (BSS) can cope also with directional nonstationary interfering sources [3]. There are also methods based on low-rank models of the RTF that can be learned in noise-free conditions. This class embodies, e.g., an approach utilizing bank of pre-learned RTFs [1] or a model based on diffusive maps [4].

Recently, a possibility to estimate the RTF using its *incomplete measurement* was studied in [5, 6]. The incomplete RTF is an RTF estimate whose values are known only for some frequencies. The estimate is completed (reconstructed) through finding its sparsest representation in the time domain. The motivation for the latter step is that typical ReIRs are fast decaying sequences, thus, appear to be compressible (approximately sparse).

In [5], the reconstruction is done through solving a weighted LASSO optimization problem. However, the optimum choice of weights is highly nontrivial,

[3] The right channel X_{R} as well as H are typically delayed by a few samples due to possible acausality of H_{RTF}. We omit this detail here for the sake of simplicity of the notation.

so only a heuristic choice is proposed. In this paper, we propose to use a different formulation based on second-order cone programming. Parameters of this formulation have clear meaning: Each parameter limits the distance of the reconstructed RTF value from its initial estimate.

2 Relative Transfer Function Estimators

Conventional frequency-domain estimator It follows from (2) that during intervals where noise signals are not active ($Y_\mathrm{L} = Y_\mathrm{R} = 0$), it is possible to estimate the RTF as

$$\widehat{H}_\mathrm{FD}(k) = \frac{\hat{\mathit{\Phi}}_{X_\mathrm{R}X_\mathrm{L}}(k)}{\hat{\mathit{\Phi}}_{X_\mathrm{L}X_\mathrm{L}}(k)}. \tag{4}$$

$\hat{\mathit{\Phi}}_{A\,B}$ denotes the sample-based estimate of (cross-)power spectral density between A and B. When signals are contaminated by noise, the estimator becomes biased where the bias (as well as its variance) depends on noise characteristics. This estimator will be abbreviated by FD (Frequency Domain estimator).

The authors of [2] considered the model where the target signal is wide-sense stationary (WSS) during short intervals (subintervals) but nonstationary over longer segments (piecewise stationary). The assumption about the noise is such that $V(k,\ell) = Y_\mathrm{R}(k,\ell) - H_\mathrm{RTF}(k)Y_\mathrm{L}(k,\ell)$ is WSS. Under this model, it was derived that the bias[4] of FD is

$$\mathrm{E}[\widehat{H}_\mathrm{FD}(k)] - H_\mathrm{RTF}(k) = \frac{\mathit{\Phi}_{VX_\mathrm{L}}(k)}{\langle \mathit{\Phi}^p_{X_\mathrm{L}X_\mathrm{L}}(k)\rangle} \tag{5}$$

where $\mathrm{E}[\cdot]$ stands for the expectation operator, and $\langle\cdot\rangle$ denotes the average of the argument over the subintervals indexed by the superscript p, $p = 1,\dots,P$. Note that the model assumes that $\mathit{\Phi}_{VX_\mathrm{L}}(k)$ is independent of p. To estimate the bias, the cross-spectral densities on the right-hand side of (5) can be replaced by their sample-based estimates; $V(k,\ell)$ can be estimated as $-Z(k,\ell)$ from (3).

Estimator admitting presence of stationary noise An estimator that is unbiased under the validity of the above model can be computed as the least-square solution of the following overdetermined set of equations [2]

$$\begin{bmatrix} \hat{\mathit{\Phi}}^1_{X_\mathrm{R}X_\mathrm{L}}(k) \\ \vdots \\ \hat{\mathit{\Phi}}^P_{X_\mathrm{R}X_\mathrm{L}}(k) \end{bmatrix} = \begin{bmatrix} \hat{\mathit{\Phi}}^1_{X_\mathrm{L}X_\mathrm{L}}(k) & 1 \\ \vdots & \\ \hat{\mathit{\Phi}}^P_{X_\mathrm{L}X_\mathrm{L}}(k) & 1 \end{bmatrix} \begin{bmatrix} \widehat{H}_\mathrm{NSFD}(k) \\ \hat{\mathit{\Phi}}_{V\,X_\mathrm{L}}(k) \end{bmatrix}. \tag{6}$$

The variance of this estimator, from here referred to as NSFD (Non-Stationarity based Frequency Domain estimator), is equal to

$$\mathrm{var}[\widehat{H}_\mathrm{NSFD}(k)] = \frac{1}{N}\frac{\mathit{\Phi}_{VV}(k)\langle 1/\mathit{\Phi}^p_{X_\mathrm{L}X_\mathrm{L}}(k)\rangle}{\langle \mathit{\Phi}^p_{X_\mathrm{L}X_\mathrm{L}}(k)\rangle\langle 1/\mathit{\Phi}^p_{X_\mathrm{L}X_\mathrm{L}}(k)\rangle - 1}. \tag{7}$$

[4] The variance of FD under the model is also derived in [2] and could be taken into account. The bias, however, seems to have a larger influence on the entire accuracy of FD; we therefore focus on the bias.

Note that $\langle \Phi^p_{X_\mathrm{L} X_\mathrm{L}}(k)\rangle\langle 1/\Phi^p_{X_\mathrm{L} X_\mathrm{L}}(k)\rangle$ is close to 1 when $\Phi^p_{X_\mathrm{L} X_\mathrm{L}}(k)$ does not depend much on p, which happens when the target signal is almost stationary (as well as the noise). By contrast, $\langle \Phi^p_{X_\mathrm{L} X_\mathrm{L}}(k)\rangle\langle 1/\Phi^p_{X_\mathrm{L} X_\mathrm{L}}(k)\rangle \gg 1$ when the signal is sufficiently dynamical. Therefore, NSFD is suitable in situations when the target signal is speech and the noise is (approximately) stationary.

It is worth to point here to a problem that is unintentionally hidden in the analysis. Speech signals are sparse in the time-frequency domain. It thus often happens that $S_\mathrm{L}(k,\ell) = 0$ for some k, which means that $H_\mathrm{RTF}(k)$ vanishes from the model (2). The behavior of NSFD then depends on the character of the (stationary) noise source. If the noise is diffused, the variance (7) approaches infinity, so we are aware of the inaccuracy of the estimate for the given frequency. However, if the noise comes from a spatial source, NSFD yields an estimate of the RTF which is related to the noise source, not to the target source.

It is important to avoid the latter case. Otherwise a large error is introduced into the estimator although (7) need not signalize it. If this case is detected through some additional hypothesis (e.g., by means of a voice-activity detector), the RTF estimate for the given k can be dropped and replaced using the method proposed in this paper. In experiments, we will focus on the described situation by considering speech as the target signal interfered by a directional quasi-stationary noise.

There are many other RTF estimators that can be taken into account in the following considerations; see, e.g., [7, 8]. Nevertheless, we will constrain our focus on the estimators FD a NSFD in this paper.

3 Sparse Reconstruction of Incomplete RTF

As already mentioned, an incomplete RTF is obtained by taking values of an RTF estimate but only for those frequencies where the estimate appears to be accurate enough. Let the set of the accepted frequencies $\{i_1, \dots, i_{|\mathcal{S}|}\}$ be denoted by $\mathcal{S}$; we can constrain $|\mathcal{S}| \le \lceil M/2+1 \rceil$ due to the symmetry of the DFT and due to the fact that the ReIR is real-valued.

The method in [5] aims to find the sparsest representation of the incomplete RTF in the time domain using weighted LASSO. The reconstructed ReIR is sought as the solution of

$$\mathbf{h}_\mathrm{WLASSO} = \arg\min_{\mathbf{h}} \|\mathbf{F}_\mathcal{S}\mathbf{h} - \mathbf{f}\|_2^2 + \|\mathbf{w} \odot \mathbf{h}\|_1, \tag{8}$$

where $\mathbf{f}$ is a $|\mathcal{S}| \times 1$ vector with elements $f_k = \widehat{H}(i_k)$, $i_k \in \mathcal{S}$, $k = 1, \dots, |\mathcal{S}|$; $\mathbf{F}$ is the $M \times M$ matrix of the DFT and $\mathbf{F}_\mathcal{S}$ denotes its submatrix comprised of rows whose indices are in $\mathcal{S}$; $\mathbf{h}$ denotes an $M \times 1$ vector of coefficients of the estimate of h_rel; $\mathbf{w}$ is an $M \times 1$ vector of nonnegative weights; $\odot$ denotes the element-wise product.

The weights control the sparsity level of the solution. They can incorporate a priori knowledge, because elements of $\mathbf{h}_\mathrm{WLASSO}$ with higher weights tend to be closer to or equal to zero and vice versa. A heuristic selection respecting the expected shape of h_rel was proposed in [5]; similar idea is used in [10].

A drawback of (8) is that the influence of the weights on the quality of the reconstructed RTF (ReIR) is not clear. In this paper, we therefore consider a different formulation where the reconstructed ReIR is defined as the solution of

$$\mathbf{h}_{\mathrm{SOCP}} = \arg\min_{\mathbf{h}} \|\mathbf{h}\|_1 \qquad \text{w.r.t.} \qquad |(\mathbf{F}_{\mathcal{S}}\mathbf{h} - \mathbf{f})_{i_k}| \leq \epsilon_{i_k}, \quad \forall i_k \in \mathcal{S}, \tag{9}$$

which is a second-order cone program (SOCP). In this formulation, the distance of the i_kth element of the reconstructed RTF from $\widehat{H}(i_k)$ is constrained to be less or equal to ϵ_{i_k} (in absolute value). For example, it is reasonable to choose ϵ_{i_k} proportional to a theoretical bias or variance of the estimate $\widehat{H}(i_k)$.

Practical Implementation Assume a stereo noisy recording obeying (2) is given. Let M be the length of DFT, which corresponds to the length of the to be estimated ReIR; for simplicity, let M be even. The proposed estimation procedure consists of four steps.

1. Compute the initial RTF estimate $\widehat{H}(k)$, $k = 0, \ldots, M/2 + 1$, using some known method. In this paper, we will consider FD given by (4) and NSFD computed through (6).
2. Compute a theoretical estimation error of $\widehat{H}(k)$, $k = 0, \ldots, M - 1$, denoted as δ_k. Here, we compute the theoretical bias (5) in case of FD and the theoretical variance (7) in case of NSFD.
3. Select $\mathcal{S}$. In this paper, we select p percents of frequency bins that yield the highest SNR (oracle selection) or the highest normalized kurtosis (kurtosis-based selection)[5]. The parameter p will be referred to as *percentage*.
4. Solve the SOCP given by (9) where $\epsilon_{i_k} = \alpha\delta_{i_k}$, $i_k \in \mathcal{S}$, using the ECOS package [9] to obtain the reconstructed ReIR; its DFT gives the reconstructed RTF; α is a free positive constant (we select $\alpha = 1$ in case of NSFD and $\alpha = 0.2$ in case of FD).

4 Experiments

In experiments, the above proposed procedures to estimate the RTF from noisy recordings are verified on real-world audio signal mixtures. A female utterance from SiSEC 2013[6] from the task "Two-channel mixtures of speech and real-world background noise" is used as the target signal. The signal has 10 s in length; the sampling frequency is 16 kHz.

The noise signal is a fan hum "FanRear.wav" by user Otakua taken from the repository of free audio samples[7]. Note that this signal is approximately

[5] The kurtosis-based selection appears to be efficient when the frequency components of the target signal have non-Gaussian distribution while those of the noise are Gaussian; see Section V in [5]. In real-world situations, this is often satisfied when the target signal is speech and the noise is quasi-stationary.

[6] `http://sisec.wiki.irisa.fr/`

[7] `http://www.freesound.org/`

stationary (as assumed by NSFD) as well as directional (spatial source). The densities of its spectral components are close to Gaussian, hence their kurtosis is close to zero. By contrast, the kurtosis of active spectral components of speech is often positive. This enables to utilize the kurtosis as a contrast to select $\mathcal{S}$.

To simulate spatial sources, the target and noise signal are convolved with room impulse responses from [11][8]. The reverberation time T_{60} is 160 ms; the distance of the microphones is 3 cm; the source-microphone distance is 2 m. The target and noise source is located, respectively, in the direction of 0° and 75° on the left-hand side.

The spatial images of the signals are mixed together at a specified input SNR (averaged over both microphones). The mixed signal is divided into 1s-intervals with 75% overlap, and the RTF estimation is conducted independently on each of total 37 intervals. The results are then given in the form of ATR averaged over all intervals.

Results and discussions Fig. 1 and Fig. 2 summarize results of the experiments with NSFD and FD, respectively. The figures show the average ATR of the estimated RTFs as functions of percentage p and of input SNR.

For both initial estimates, the RTFs reconstructed by SOCP yield ATR that is comparable or higher to that obtained by weighted LASSO. This holds for both the oracle and kurtosis-based selection of $\mathcal{S}$. The improvements compared to LASSO are achieved for low values of p (below $10-20\%$) and for scenarios with the lower input SNR (see Fig. 1(a,c) and Fig. 2(a)). When input SNR is -10 dB, only few frequency components of the initial RTF estimate are accurate “enough”, so small p should be selected. Then, SOCP appears to be more robust than LASSO. For higher values of p and higher input SNR, both approaches achieve comparable ATRs (see Fig. 1(b,d) and Fig. 2(b)).

The overall ATRs with FD are significantly lower than those with NSFD; cf. Fig. 1(a) and Fig. 2(a). This confirms the assumption that NSFD yields more accurate RTF estimate, when target speech is interfered by a stationary noise.

Fig. 1(c,d) and Fig. 2(b) show that the reconstructed RTFs from incomplete measurements yield significant improvement in terms of ATR, especially, when input SNR is low. When input SNR is sufficiently high, the ATR by the reconstructed RTFs can be lower than that of the initial estimates, depending on p. When p is too low, the loss of the ATR signalizes that too much information was lost from the incomplete RTF (see Fig. 1(c) where input SNR>0). By contrast, when the RTF measurement is almost complete (p close to 100%), the ATRs by the reconstructed RTFs are getting closer to those of the initial estimators.

The oracle selection yields higher ATR compared to kurtosis-based selection of $\widehat{H}(k)$ for all values of p and all considered input SNR levels (by up to 3 dB). This is due to the strong prior knowledge utilized by the oracle selection (the true input SNR within the frequency bins).

[8] `http://www.eng.biu.ac.il/gannot/downloads/`

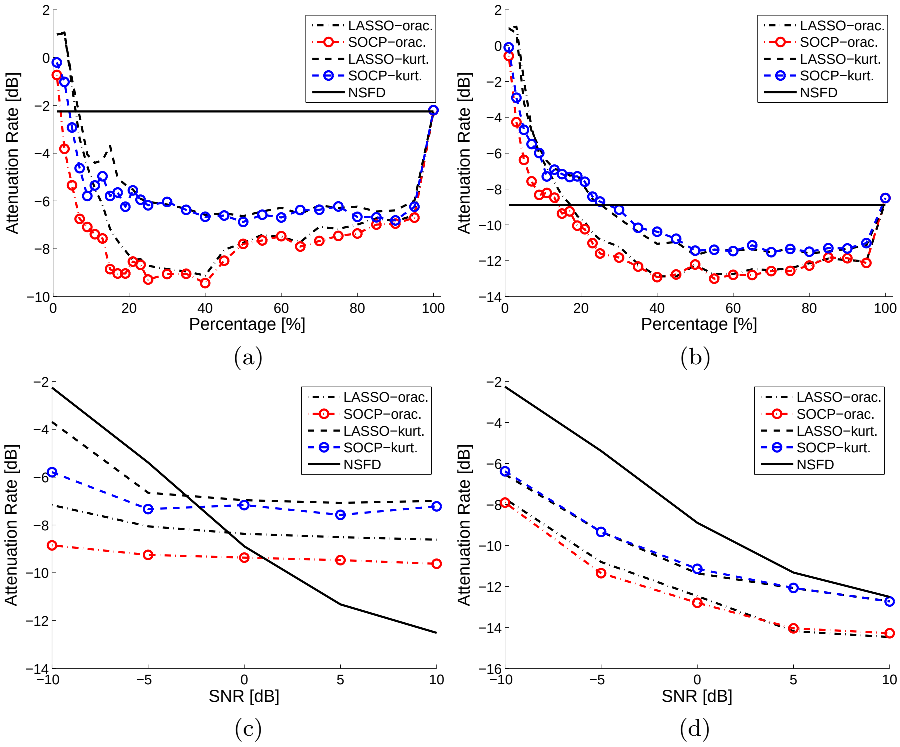

Fig. 1. Attenuation of female speech in the presence of directional fan hum. The initial estimate $\widehat{H}(k)$ is NSFD (6). (a,b) Dependence on the percentage of frequency bins included in $\mathcal{S}$ (input SNR = -10 dB or 0 dB, respectively), (c,d) dependence on input SNR ($p = 15\%$ or $p = 65\%$,respectively). The more negative the value (in dBs) of ATR is, the better the target signal blocking.

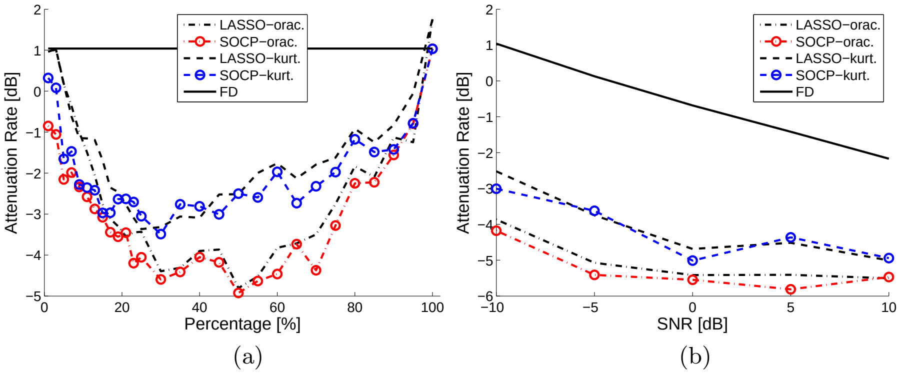

Fig. 2. Attenuation of female speech in the presence of directional fan hum. The initial estimate $\widehat{H}(k)$ is FD (4). (a) Dependence on the percentage of frequency bins included in $\mathcal{S}$ (input SNR = -10 dB), (b) the dependance on input SNR ($p = 45\%$).

5 Conclusions

We observed that the solutions by LASSO and SOCP can be very close in the sense that an appropriate choice of weights in LASSO enables to approach the solution by SOCP. However, the correspondence between the parameters of LASSO and SOCP is nontrivial. In contrast to the weighted LASSO, the interpretation of the parameters in SOCP is straightforward and helpful in practice, which was demonstrated by experiments in this paper.

References

1. 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.
2. Shalvi, O., Weinstein, E., “System Identification Using Nonstationary Signals,” IEEE Trans. Signal Processing, 44(8):2055–2063, Aug. 1996.
3. Parra, L., and Spence, C.: “Convolutive Blind Separation of Non-Stationary Sources”, *IEEE Trans. on Speech and Audio Processing*, 8(3):320–327, May 2000.
4. Talmon, R., Gannot, S., “Relative Transfer Function Identification on Manifolds for Supervised GSC Beamformers,” Proc. of the 21st European Signal Processing Conference (EUSIPCO), pp. 1–5, Marrakech, Morocco, Sep. 2013.
5. Koldovský, Z., Málek, J., and Gannot, S., “Spatial Source Subtraction Based on Incomplete Measurements of Relative Transfer Function,” *IEEE Trans. on Speech, Audio and Language Processing*, 2015.
6. 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.
7. Takahashi, Y., Takatani, T., Osako, K., Saruwatari, H., Shikano, K., “Blind Spatial Subtraction Array for Speech Enhancement in Noisy Environment,” IEEE Transactions on Audio, Speech, and Language Processing, 17(4):650–664, May 2009.
8. Nesta, F., Matassoni, M., Astudillo, R.F., “A Flexible Spatial Blind Source Extraction Framework for Robust Speech Recognition in Noisy Environments,” Proc. of the 2nd CHiME Workshop on Machine Listening in Multisource Environment, pp. 33–40, June 2013.
9. Domahidi, A., Chu, E., Boyd, S., “ECOS: An SOCP Solver for Embedded Systems,” Proceedings of European Control Conference, pp. 3071–3076, Zurich, July 2013.
10. Benichoux, A., Simon, L. S. R., Vincent, E., and Gribonval, R., “Convex Regularizations for the Simultaneous Recording of Room Impulse Responses,” IEEE Transactions on Signal Processing, 62(8):1976–1986, April 2014.
11. E. Hadad, F. Heese, P. Vary, and S. Gannot, “Multichannel audio database in various acoustic environments,” *International Workshop on Acoustic Signal Enhancement 2014 (IWAENC 2014)*, pp. 313–317, Antibes, France, Sept. 2014.