ESTIMATION OF TDOA FOR ROOM REFLECTIONS BY ITERATIVE WEIGHTED L1 CONSTRAINT

Marco Crocco and Alessio Del Bue

Pattern Analysis and Computer Vision (PAVIS)
Istituto Italiano di Tecnologia (IIT)
Via Morego 30, 16163 Genova, Italy

ABSTRACT

Estimation of Time Difference of Arrivals (TDOAs) corresponding to early room reflections can be formulated as a Blind Channel Identification (BCI) problem, exploiting signals acquired by a set of microphones, given an unknown transmitted source. To cope with the intrinsic noise sensitivity and ill-posedness of the problem, sparsity and non-negativity priors on the Acoustic Impulse Response (AIR) of the room can be exploited. Here we propose a novel iterative method, resulting in a sequence of convex problems relying on a weighted l_1 constraint. The proposed method allows to outperform current state of the art on speech and non-speech real signals, while lowering the number of parameters to tune and the sensitivity of solution given the few free parameters.

Index Terms— Acoustic Impulse Response, Blind Channel Identification, Sparsity, Non-negative Priors, TDOA Estimation

I. INTRODUCTION

The robust inference of TDOAs in a room is a fundamental step in several audio processing applications, such as room aware sound reproduction [1], inference of room geometry [2], [3], [4], [5], [6], speech enhancement [7] and dereverberation [8], [9]. A common setup is given by a set of microphones deployed in the room and a audio source, whose transmitted signal is typically unknown (natural signals). Baseline methods based on cross-correlation among acquired signals are poorly performing, especially with noisy environments, transmitted signals of limited bandwidth and low energy reflections with respect to the direct path [10]. An attractive alternative consists in blindly estimating the room AIRs and extracting TDOAs from the AIRs peaks. This approach results in a Single Input Multi Output (SIMO) blind channel identification problem that can be solved by exploiting the spatial diversity of channels related to each microphone couple [11].

Following works [12], [13], [14], [15] have exploited prior knowledge on the AIRs in order to relax the strict feasibility conditions of [11] and provide more robustness toward environmental noise. In this paper we propose a novel iterative optimization strategy aimed at further improving the accuracy and robustness of AIR estimation. The key features of our method are the elimination of strong constraints leading to biased solutions (e.g. the anchor constraint [12], [13]), the reduction of free parameters that ease the tuning of the method, and the ability to promote sparsity and non-negativity to increase robustness and accuracy. Moreover, each iteration of the proposed method results in a convex problem that can be easily solved with quadratic programming techniques. The experimental tests based on synthetic and real acoustic signals (both speech and non-speech) compare favourably to our approach, always achieving superior performance in respect to the state of the art.

The next section formalises the AIR estimation problem together with previous approaches contributions. Sec. III describes our method formalisation and optimization strategy. Experimental results in Sec. IV show the performance of several methods compared to ours while in Sec. V some conclusions are drawn.

II. PROBLEM STATEMENT

Let us consider N microphones in a room and let us define $h_n(k)$ as the discrete time AIR from a source to the n-th microphone. The signal $y_n(k)$ received at microphone n can be written as the discrete convolution between the transmitted signal $x(k)$ and the n-th AIR:

$$y_n(k) = h_n(k) * x(k) + \nu_n(k), \qquad n = 1, \dots, N \tag{1}$$

where $\nu_n(k)$ is an additive noise term. The BCI problem aims at recovering the set of AIRs $h_n(k)$ without knowing the transmitted signal $x(k)$. A family of methods is based on the cross-relation identity for which, in absence of noise, the equality $y_n(k) * h_m(k) = y_m(k) * h_n(k)$ holds for every couple (n, m). This principle is used in [11] by introducing a Least Squares minimization of the squared cross relation error as:

$$\min_{\mathbf{h}} \sum_{m \neq n} \|\mathtt{Y}_n \mathbf{h}_m - \mathtt{Y}_m \mathbf{h}_n\|_2^2 \quad s.t. \quad \|\mathbf{h}\|^2 = 1, \tag{2}$$

where $\mathbf{h}_n = [h_n(1), \cdots, h_n(L)]^\top$ and $\mathbf{h} = [\mathbf{h}_1^\top, \dots \mathbf{h}_N^\top]^\top$. The matrix $\mathtt{Y}_n$ is Toeplitz with first row and column given by $[y_n(k-K+1), y_n(k-K), \dots, y_n(k-K-L+2)]$ and $[y_n(k-$

$K+1), y_n(k-K+2), \ldots, y_n(k), 0, \ldots, 0]^\top$ respectively, with K and L being the signal length and channel length respectively, and . The equality constraint on the l_2-norm of $\mathbf{h}$ avoids the trivial zero solution. By rearranging the matrices $\mathtt{Y}_n$, the above minimization problem can be solved with an eigenvalue decomposition [11].

However, the recovery of channels $\mathbf{h}$ from Eq. (2) is dependent from a set of restrictive assumptions. In particular, the channels have to be co-prime, hence their length L has to be known in advance. This information is not available in real situations [11] and if the channel length is not correct, the problem is ill-conditioned and therefore highly sensitive to environmental noise. Moreover, if the spectrum of $x(k)$ contains "holes", a likely case with real signals, some frequency components of $\mathbf{h}$ in Eq. (2) will be weighted by zero or very small values. This has a subtle but disruptive effect, since the energy of $\mathbf{h}$, constrained by the l_2-norm, may be forced to concentrate at these frequencies contributing little in the cost function, so leading to grossly distorted solutions.

This method has been subsequently improved exploiting the a priori knowledge of the AIR structure [16]. In fact, the first part of the room AIR can be modelled as a set of positive pulses, each one given by the direct path or a reflection from a wall [16]. Thus, sparsity [12], [14] and non-negativity [13] have been imposed to improve robustness. Even if sparsity is verified only for early TDOAs and this is not valid for the 'tail" of the AIR, applications on room reconstruction only leverage on the sparse component [2], [3], [4], [5], [6], [1], [15], thus it is extremely important to design methods able to extract this information. Moreover, speech enhancement [17] and derevereberation [8] and room aware sound reproduction have proven to work assuming AIR sparsity.

In [14] a l_1-norm penalty was added to Eq. (2) in order to enforce sparsity, yielding:

$$\min_{\mathbf{h}} \; J(\mathbf{h}) + \lambda \|\mathbf{h}\|_1 \quad s.t. \quad \|\mathbf{h}\|_2^2 = 1, \tag{3}$$

with $J(\mathbf{h}) = \sum_{n \neq m} \|\mathtt{Y}_m \mathbf{h}_n - \mathtt{Y}_n \mathbf{h}_m\|_2^2$. Unfortunately, this regularization introduces a fundamental drawback: the domain of the problem is non-convex due to the quadratic equality constraint, so making the minimization of $J(\mathbf{h}) + \lambda\|\mathbf{h}\|_1$ prone to local solutions.

To cope with this issue, an anchor constraint can be used to replace the l_2-norm one giving [12]:

$$\min_{\mathbf{h}} \; J(\mathbf{h}) + \lambda \|\mathbf{h}\|_1 \quad s.t. \quad |h_1(a)| = 1, \tag{4}$$

where a is the anchor index [1]. The anchor constraint makes the problem convex [18] and less sensitive to spectrum holes than l_2-norm constraint [12]. However, the anchor constraints together with the l_1-norm penalizes all the peaks intensities but one, often leading to peak cancellations in noisy conditions. The approach of [12] has been modified in [13] adding the non-negativity constraint on the AIR $\mathbf{h} \geq 0$ (i.e. $h_n(k) \geq 0$ for $k = 1, \ldots, L$ and $n = 1, \ldots, N$). Non-negativity yields increased robustness against noise by further regularizing the problem [19], [20].

[1]The anchor index has to be chosen greater than the maximum of the differences $\tilde{k}_n - \tilde{k}_1$ over $n = 1, \cdots, N$ where $\tilde{k}_n$ is the index of the first non zero entry of the channel n.

III. PROPOSED METHOD

To solve the drawbacks related to the anchor constraint in Eq. (4), we introduce an l_1-norm equality constraint, obtaining:

$$\min_{\mathbf{h}} J(\mathbf{h}) + \lambda \|\mathbf{h}\|_1 \quad s.t. \quad \|\mathbf{h}\|_1 = 0, \quad \mathbf{h} \geq 0. \tag{5}$$

In this way, all the channel elements are equally taken into account without privileging the one corresponding to the anchor. At the same time, differently from the l_2-norm constraint, the problem remains convex and differentiable as, due to the non-negativity constraint, the l_1-norm becomes a simple sum of the elements of $\mathbf{h}$. But setting such l_1 constraint destroys the sparsity-inducing effect of the l_1 penalty. In fact the l_1 penalty becomes a constant that does not influence the argument of the minimum of the cost function.

To tackle this issue we propose to solve a sequence of minimization problems of the form:

$$\hat{\mathbf{h}}^{(z)} = \min_{\mathbf{h}} J(\mathbf{h}) + \lambda \|\mathbf{h}\|_1 \quad s.t. \quad \mathbf{p}^{(z)\top}\mathbf{h} = 1, \;\; \mathbf{h} \geq 0, \tag{6}$$

where $\mathbf{p}^{(z)}$ is a $L \times 1$ weight vector. At each iteration, $\mathbf{p}^{(z)}$ is made equal to the solution of the problem at the previous step $(z-1)$:

$$\mathbf{p}^{(z)} = \hat{\mathbf{h}}^{(z-1)}. \tag{7}$$

In practice we substitute the standard l_1-norm constraint with an adaptive weighted l_1-norm constraint $\mathbf{p}^{(z)\top}\mathbf{h} = 1$.

Concerning the algorithm initialization $\hat{\mathbf{h}}^{(0)}$, we adopt the solution provided by [13], i.e. the standard l_1 penalty with anchor and non-negativity constraints. Such initialization generally assures a sufficient degree of sparsity for the starting guess. In the subsequent steps, the weighted constraint $\mathbf{p}^{(z)\top}\mathbf{h} = 1$ will further enforce sparser solutions. If the equality constraint gives more importance to larger elements of the channels, the unweighted l_1-norm penalty in the cost function will be able to penalize selectively the smaller elements, as can be seen comparing iterations 0 and 1 in Fig. 1. Also notice that the initial amplitude distortion introduced by the anchor is compensated by the subsequent steps of the algorithm (see Fig. 1).

Geometrical Interpretation. In order to gain a deeper insight on the effects of the weight vector $\mathbf{p}$[2], let us introduce a new variable $\mathbf{k}$ defined as: $\mathbf{k} = \mathbf{p} \odot \mathbf{h}$, where $\odot$ denotes

[2]Index (z) dropped for simplicity.

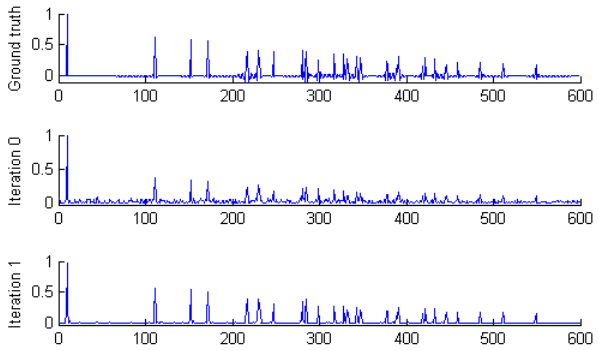

Fig. 1. Example of the iterative process for our proposed approach with SNR = 14 dB

the Hadamard product. Substituting $\mathbf{h}$ with $\mathbf{k}$ in Eq. (6), we can reformulate the problem as:

$$\|\mathtt{XWk}\|_2^2 + \lambda \mathbf{w}^\top \mathbf{k} \quad s.t. \quad \|\mathbf{k}\|_1 = 1 \quad \mathbf{k} > 0, \qquad (8)$$

where $\mathbf{w}$ is a new weight vector whose elements are the inverse of the elements of $\mathbf{p}$, $\mathtt{X}$ is a reorganization of matrices $\mathtt{X}_i$ such that $J(\mathbf{h}) = \|\mathtt{Xh}\|_2^2$ and the matrix $\mathtt{W}$ is given by $\mathtt{W} = diag(\mathbf{w})$ where $diag$ denotes the operator that maps a vector into a diagonal matrix. Now, let us give a geometric interpretation of the quadratic and the linear term of the cost function in Eq. (8) given the equality constraint $\|\mathbf{k}\|_1 = 1$. The term $\mathbf{w}^\top \mathbf{k}$ is equal to a weighted l_1 norm i.e. a "weighted l_1 polyhedron" centered at the origin and "pinched" toward the directions corresponding to the smallest elements of $\mathbf{w}$. On the other hand, the constraint $\|\mathbf{k}\|_1 = 1$ is a uniform l_1 ball of radius 1. It is clear that the minimization of $\mathbf{w}^\top \mathbf{k}$ on the l_1 ball $\|\mathbf{k}\|_1 = 1$ is satisfied at the vertices corresponding to the smallest elements of $\mathbf{w}$, thus enforcing the sparsity of the solution.

Instead, the quadratic term can be rewritten as $\|\mathtt{XWk}\|_2^2 = \mathbf{k}^\top \mathtt{W}\hat{\mathtt{X}}\mathtt{W}\mathbf{k}$, where $\hat{\mathtt{X}} = (\mathtt{X}^\top \mathtt{X})$. The positive semi-definite matrix $\hat{\mathtt{X}}$ defines a multidimensional ellipsoid centred on the coordinates origin. The pre- and post-multiplication by the diagonal matrix $\mathtt{W}$ acts as an affine operator that shrinks and leans the ellipsoid axes in such a way that the new ellipsoid $\mathtt{W}\hat{\mathtt{X}}\mathtt{W}$ will tend to align its larger axes along the coordinates corresponding to the smallest elements of $\mathbf{w}$ i.e. the largest elements of $\hat{\mathbf{h}}^{(z-1)}$. This in turn will enforce a solution $\hat{\mathbf{h}}^{(z)}$ in which larger components are even larger and the smaller ones are even smaller, thus promoting a sparser solution. Therefore, the proposed method induces the sparsity of the solution by a twofold mechanism, while solving at the same time the issue of the drawbacks of the anchor constraints. A recent method [15] follows a somewhat similar strategy,

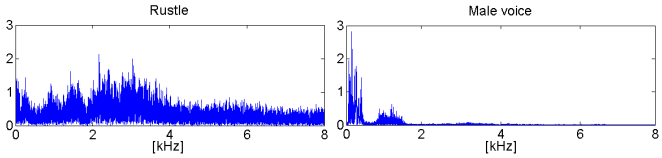

Fig. 2. Spectra of non-speech (left) and speech (right) recorded signal used in the experiments.

solving a sequence of problems:

$$\hat{\mathbf{h}}^{(z)} = \min_{\mathbf{h}} \|\mathtt{Xh}\|_2^2 + \lambda \tilde{\mathbf{w}}^{(z)\top} \mathbf{h} \quad s.t. \quad \|\mathbf{h}\|_1 = 1, \quad \mathbf{h} \geq 0, \qquad (9)$$

with $\tilde{w}(k)^{(z)} = 1/\left[\hat{h}(k)^{(z-1)} + \epsilon\right]$ for $k = 1, \cdots, NL$, where $\tilde{w}(k)^{(z)}$ and $\hat{h}(k)^{(z-1)}$ denote the k-th element of $\tilde{\mathbf{w}}^{(z)}$ and $\hat{\mathbf{h}}^{(z-1)}$ respectively and ϵ is a regularization parameter. Here just the l_1 penalty is weighted, while the quadratic part $\|\mathtt{Xh}\|_2^2$ is left unchanged, so limiting the sparsity-inducing effect. Moreover, since $\tilde{\mathbf{w}}$ is explicitly calculated, it is necessary to choose a value for ϵ to avoid numerical instabilities. Differently, in our approach the weights $\mathbf{w}$ are only implicitly evaluated, since in the implemented formulation just the weights $\mathbf{p}$ are calculated.

IV. EXPERIMENTS

The proposed algorithm, named here Iterative L1 Constraint (IL1C), was compared with the eigenvalue decomposition (EIG) approach [11], the non-negative l_1-norm method [13] (L1NN) and the approach of [15], named here Iterative l_1 Penalty (IL1P). We considered a rectangular room of $6 \times 5 \times 4$ m with microphones and sources position randomly generated at each trial [3]. The number of microphones was fixed to 2. The AIRs were simulated according to the image method [16] assuming a reflection coefficient of 0.8. Three sources were used: a 1 s synthetic white noise, a 2 s real recorded signal consisting in a rustle caused by plastic material and a 1.8 s male voice segment, all sampled at 16 kHz (check Fig. 2 for the related spectra). The length of the estimated channel was of $L = 700$ samples, such value being an upper bound for every possible real channel length, given the room geometry. Gaussian white noise was added to the signals at each microphone output, in order to model microphone self noise, setting the following SNR values: 20 dB, 14 dB 6 dB and 0 dB. All the parameters for all the tested methods (λ, a and ϵ) were optimized by cross-validation. For IL1C and IL1P, the number of iterations was set to 4. Fig. 3 shows an example of the real and estimated channel for the four methods, assuming a challenging SNR of 6 dB. As it can be seen, EIG completely fails due to its extreme noise sensitivity. L1NN provides a reasonable but very noisy AIR

[3]To avoid configurations in which the source is too close to the microphones, the x coordinate ranged from 0 m to 2 m for the microphones and from 4 m to 6 m for the source.

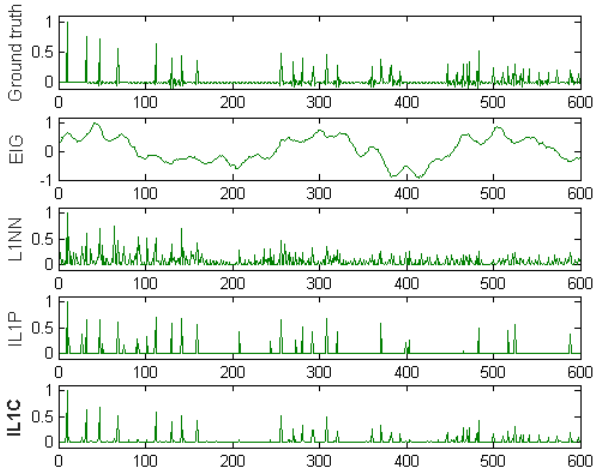

Fig. 3. Results for the four tested methods and ground truth. SNR = 6 dB. Synthetic source.

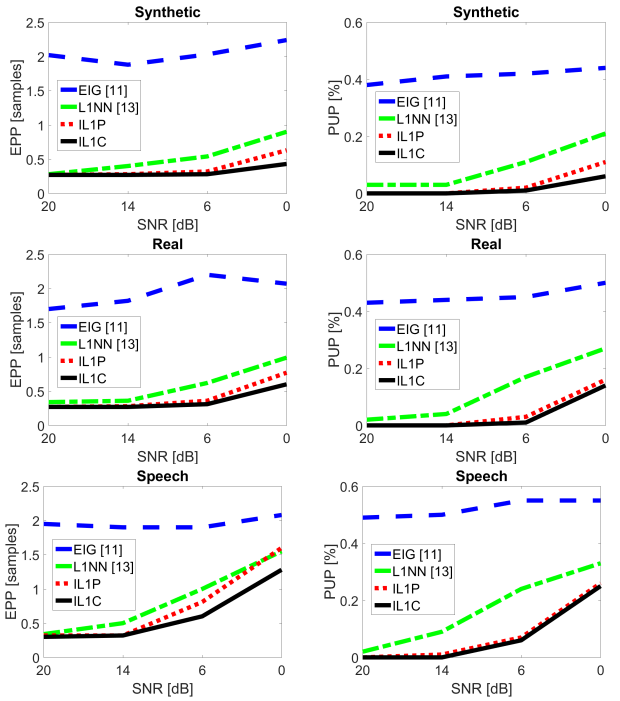

Fig. 4. EPP (left) and PUP (right) versus SNR level for synthetic, real and speech signals.

reconstruction, from which TDOAs are hardly detectable. An improved solution is achieved with IL1P but several spurious peaks are present and a number of true peaks is canceled out. Differently, a very accurate result is achieved by IL1C [4]. Quantitative results over 50 Monte Carlo simulations are reported in Fig. 4 for synthetic, non-speech and speech sources versus SNR. The average error in samples of estimated peaks position (EPP), limited to matched peaks, and the percentage of unmatched peaks (PUP) are displayed in left and right panels respectively. A Ground Truth (GT) peak is considered unmatched if the closest estimated peak is more than 10 samples away from it. These two metrics allow to decouple the effect of the outliers, from the overall peak position accuracy. The matching between GT peaks and estimated ones is performed, for each channel, by a Nearest Neighbour procedure repeated for each ground truth peak[5].

As can be seen from Fig. 4, EIG yields very bad performance, almost independent from the SNR level. L1NN manages to obtain reasonable results but the percentage of outliers and the average precision are not satisfying for high SNR values. IL1P improves on average the results of L1NN while the proposed methods IL1C outperforms all the other ones. In particular IL1C performance is equal to IL1P in a few cases, for high SNR values, but it is significantly higher when the SNR decreases. Moreover, IL1C performance is less dependent on the parameter λ than IL1P, as shown by Table I, where results are averaged over three orders of magnitude of λ. For IL1P the mean values are by far increased with respect to the values in Table 1, that were related to the best λ, while for IL1C just a moderate worsening is observed. Moving from synthetic to speech results in general get worse, likely due to the progressively decreasing flatness of the source spectrum that limits the amount of frequencies available (check Fig. 2). Nevertheless the good performance of the proposed method favours its applicability to real noisy environments.

	SNR	IL1P [15]	IL1C
SY	0 dB	1.28 [0.22]	**0.49 [0.10]**
	6 dB	0.53 [0.09]	**0.32 [0.02]**
RE	0 dB	1.49 [0.21]	**0.70 [0.16]**
	6 dB	0.46 [0.06]	**0.34 [0.03]**
SP	0 dB	2.27 [0.26]	**1.32 [0.26]**
	6 dB	1.12 [0.14]	**0.80 [0.12]**

Table I. EPP and PUP (in squared brackets) for different sources. Results are averaged over different values of λ.

V. CONCLUSIONS

The proposed method has proven to increase the accuracy of estimation of AIRs and consequently TDOAs, in respect to state of the art methods, in challenging realistic conditions. Future work will test the benefits of the method when applied to tasks such as room geometry reconstruction.

[4]Note that the real channel is not perfectly non-negative since the Dirac pulses, corresponding to the walls reflections, typically fall off the grid of samples resulting in sinc functions. Despite this slight mismatch between theoretical assumptions and real data, the position of the estimated peaks reproduces the positions of the ground truth peaks with remarkable precision.

[5]To avoid results biased in favour of solutions yielding many peaks, the K_c estimated peaks with the highest amplitudes have been taken into account, where K_c is the number of ground truth peaks for each channel. Peaks are found by a robust peak finder: http://www.mathworks.com/matlabcentral/fileexchange/25500-peakfinder based on local slope features.

VI. REFERENCES

[1] T. Betlehem and T. D. Abhayapala, "A modal approach to soundfield reproduction in reverberant rooms," in *2005 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP '05, Philadelphia, Pennsylvania, USA, March 18-23, 2005*, 2005, pp. 289–292.

[2] I. Dokmanić, R. Parhizkar, A. Walther, Y. M. Lu, and M. Vetterli, "Acoustic echoes reveal room shape," *Proceedings of the National Academy of Sciences*, vol. 110, no. 30, pp. 12 186–12 191, 2013.

[3] F. Antonacci, J. Filos, M. R. Thomas, E. A. P. Habets, A. Sarti, P. A. Naylor, and S. Tubaro, "Inference of room geometry from acoustic impulse responses," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 10, pp. 2683–2695, 2012.

[4] F. Ribeiro, D. Florêncio, D. Ba, and C. Zhang, "Geometrically constrained room modeling with compact microphone arrays," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 5, pp. 1449–1460, 2012.

[5] A. H. Moore, M. Brookes, and P. A. Naylor, "Room geometry estimation from a single channel acoustic impulse response," in *Signal Processing Conference (EUSIPCO), 2013 Proceedings of the 21st European*. IEEE, 2013, pp. 1–5.

[6] M. Crocco, , A. Trucco, V. Murino, and A. Del Bue, "Towards fully uncalibrated room reconstruction with sound," in *22nd European Signal Processing Conference (EUSIPCO), Lisbon, Portugal*, 2014.

[7] M. Wu and D. Wang, "A two-stage algorithm for one-microphone reverberant speech enhancement," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 3, pp. 774–784, 2006.

[8] Y. Lin, J. Chen, Y. Kim, and D. D. Lee, "Blind channel identification for speech dereverberation using l1-norm sparse learning," in *Advances in Neural Information Processing Systems*, 2007, pp. 921–928.

[9] K. Lebart, J.-M. Boucher, and P. Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acta Acustica united with Acustica*, vol. 87, no. 3, pp. 359–366, 2001.

[10] J. Chen, J. Benesty, and Y. Huang, "Time delay estimation in room acoustic environments: An overview." *EURASIP J. Adv. Sig. Proc.*, vol. 2006, 2006.

[11] L. Tong, G. Xu, and T. Kailath, "Blind identification and equalization based on second-order statistics: a time domain approach," *Information Theory, IEEE Transactions on*, vol. 40, no. 2, pp. 340–349, Mar 1994.

[12] Y. Lin, J. Chen, Y. Kim, and D. D. Lee, "Blind channel identification for speech dereverberation using l1-norm sparse learning," in *Advances in Neural Information Processing Systems*, 2007, pp. 921–928.

[13] Y. Lin, J. Chen, Y. Kim, and D. Lee, "Blind sparse-nonnegative (bsn) channel identification for acoustic time-difference-of-arrival estimation," in *Applications of Signal Processing to Audio and Acoustics, 2007 IEEE Workshop on*, Oct 2007, pp. 106–109.

[14] K. Kowalczyk, E. Habets, W. Kellermann, and P. Naylor, "Blind system identification using sparse learning for tdoa estimation of room reflections," *Signal Processing Letters, IEEE*, vol. 20, no. 7, pp. 653–656, July 2013.

[15] M. Crocco and A. Del Bue, "Room impulse response estimation by iterative weighted l1 norm," in *23nd European Signal Processing Conference (EUSIPCO), Nice, France*, 2015.

[16] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[17] M. Yu, W. Ma, J. Xin, and S. Osher, "Multi-channel l_1 regularized convex speech enhancement model and fast computation by the split bregman method," *Audio, Speech, and Lang. Proc., IEEE Trans. on*, vol. 20, no. 2, pp. 661–675, Feb 2012.

[18] S. Boyd and L. Vandenberghe, *Convex Optimization*. New York, NY, USA: Cambridge University Press, 2004.

[19] L. Benvenuti and L. Farina, "A tutorial on the positive realization problem," *Automatic Control, IEEE Transactions on*, vol. 49, no. 5, pp. 651–664, 2004.

[20] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, 2001, pp. 556–562.