On time delay estimation from a sparse linear prediction perspective

Hongsen He, Tao Yang, and Jingdong Chen

Citation: The Journal of the Acoustical Society of America **137**, 1044 (2015); doi: 10.1121/1.4906267
View online: https://doi.org/10.1121/1.4906267
View Table of Contents: https://asa.scitation.org/toc/jas/137/2
Published by the Acoustical Society of America

On time delay estimation from a sparse linear prediction perspective (L)

Hongsen He[a] and Tao Yang
School of Information Engineering and Robot Technology Used for Special Environment Key Laboratory of Sichuan Province, Southwest University of Science and Technology, Mianyang 621010, China

Jingdong Chen
Center of Intelligent Acoustics and Immersive Communications, Northwestern Polytechnical University, 127 Youyi West Road, Xi'an 710072, China

(Received 18 October 2014; accepted 8 January 2015)

This paper proposes a sparse linear prediction based algorithm to estimate time difference of arrival. This algorithm unifies the cross correlation method without prewhitening and that with prewhitening via an ℓ_2/ℓ_1 optimization process, which is solved by an augmented Lagrangian alternating direction method. It also forms a set of time delay estimators that make a tradeoff between prewhitening and non-prewhitening through adjusting a regularization parameter. The effectiveness of the proposed algorithm is demonstrated in noisy and reverberant environments.
 [http://dx.doi.org/10.1121/1.4906267]

[KGS] Pages: 1044–1047

I. INTRODUCTION

Time delay estimation (TDE), which is used to measure the relative time difference of arrival among spatially separated sensors, plays an important role in localizing and tracking talkers in room acoustic environments. The generalized cross-correlation (GCC) method[1] is by far the most popular TDE technique, which obtains the time delay between two microphones as the time lag that maximizes the cross correlation function between filtered versions of the received signals. However, TDE performance of GCC was found to deteriorate significantly when reverberation or noise is high. Many new ideas have recently been proposed to better deal with noise and reverberation, such as the multichannel cross-correlation coefficient algorithm,[2] the multichannel spatio-temporal prediction algorithm,[3] the blind channel identification algorithm,[4,5] the information theory based algorithms,[6,7] etc. Among all of these efforts, one simple yet effective way of improving the robustness of TDE against reverberation is to incorporate a prewhitening process as used in the phase transform (PHAT) algorithm.[1,8] As far as the prewhitening is concerned, linear prediction is an important technique,[9] which has already been applied to TDE (Ref. 10) and acoustic source localization.[11] The configuration of the traditional linear predictor uses a cascade of a long-term predictor and a short-term predictor.[12] The consequent prediction coefficient vector is highly sparse.[13] However, this sparsity is reduced or even not present when speech signals are contaminated by noise, which degrades the performance of the linear predictor.

In this paper, we propose a sparse linear prediction algorithm and investigate the effect of the prewhitening with different levels on TDE performance. The new algorithm introduces a sparse regularization term to the least squares criterion to form an ℓ_2/ℓ_1 optimization method, which unifies the cross-correlation (CC) and the GCC-PHAT algorithms from a TDE performance perspective. Meanwhile, we propose to use an augmented Lagrangian alternating direction method (ADM) (Ref. 14) to solve the optimization problem of the linear predictor. The performance of the new approach is demonstrated via numerical experiments.

[a]Author to whom correspondence should be addressed. Also at: Key Laboratory of Modern Acoustics of MOE, Nanjing University, Nanjing 210093, China. Electronic mail: hongsenhe@gmail.com

II. TDE VIA SPARSE LINEAR PREDICTION

A. Algorithm derivation

Usually, one can directly compute the CC function between two microphone signals $x_1(n)$ and $x_2(n)$ for TDE. The CC approach has been shown to be robust against background noise.[4] This method, however, is sensitive to room reverberation. One way to improve the robustness of the CC method to reverberation is through the use of a prewhitening process, as in the well-known GCC-PHAT algorithm. In this paper, we consider the prewhitening from a linear prediction perspective.

We consider the prediction of the current sample of channel m ($m = 1, 2$) from its past samples, i.e.,

$$x(n) = \sum_{k=1}^{K} a_k x(n-k) + e(n), \tag{1}$$

where a_k, $k = 1, 2, \ldots, K$, are prediction coefficients, K is the length of the predictor, and $e(n)$ is the prediction error. Note that we have dropped the subscript m for the simplicity of notation. In vector/matrix form, the signal in Eq. (1) can be written as

$$\mathbf{x}(n) = \mathbf{X}(n)\mathbf{a} + \mathbf{e}(n), \tag{2}$$

where

$$\mathbf{x}(n) = [x(n), x(n+1), \ldots, x(n+K+L-1)]^T, \tag{3}$$

$$\mathbf{a} = [a_1, a_2, \ldots, a_K]^T, \tag{4}$$

$$\mathbf{e}(n) = [e(n), e(n+1), \ldots, e(n+K+L-1)]^T, \tag{5}$$

$$\mathbf{X}(n)=\begin{bmatrix} x(n-1) & x(n-2) & \cdots & x(n-K) \\ x(n) & x(n-1) & \cdots & x(n-K+1) \\ \vdots & \vdots & \ddots & \vdots \\ x(n+K+L-2) & x(n+K+L-3) & \cdots & x(n+L-1) \end{bmatrix}, \tag{6}$$

L is the frame length, and $(\cdot)^T$ denotes the transpose of a vector or matrix.

The most commonly used criterion to solve Eq. (2) is the least squares method.[15] The configuration of the conventional predictor uses a cascade of a long-term predictor and a short-term predictor. This structure is motivated from the speech production model, which decouples the quasi-periodic source (the vocal folds) from the vocal tract filter.[12] The consequent prediction coefficient vector is highly sparse,[13] as illustrated in Fig. 1(a). This sparsity, however, is greatly affected by the presence of noise, which can be seen from Fig. 1(b). Since the prediction vector is sparse for clean speech signals, we can use this property to improve the robustness of the estimation of the linear predictor in noise. To this end, we introduce an ℓ_1-norm based sparse regularization term to the least squares criterion to impose the sparsity of the prediction vector. So, we propose the following ℓ_2/ℓ_1 optimization criterion to preprocess the microphone signals:

$$\min_{\mathbf{a},\mathbf{e}(n)}\left\{\frac{1}{2}\|\mathbf{e}(n)\|_{\ell_2}^2+\lambda\|\mathbf{a}\|_{\ell_1}:\mathbf{e}(n)=\mathbf{x}(n)-\mathbf{X}(n)\mathbf{a}\right\}, \tag{7}$$

where $\|\cdot\|_{\ell_2}$ and $\|\cdot\|_{\ell_1}$ stand for the ℓ_2-norm and ℓ_1-norm, respectively, and the parameter $\lambda>0$ is a scalar regularization parameter.

It is obvious that Eq. (7) is a convex optimization problem, which can be solved by many existing methods, such as the linear programming,[16] the interior point method,[17] the primal-dual interior point method,[18] etc. Unlike the above approaches, we adopt the ADM, which efficiently uses the separability of multiple variables,[14] to solve this problem.

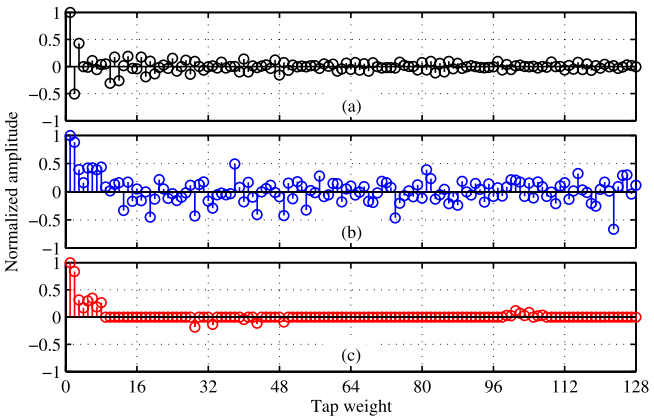

FIG. 1. (Color online) Illustration of the linear prediction vector, where the predictor length is 128, the length of the speech frame is 1024. (a) Linear prediction vector of a clean speech signal; (b) linear prediction vector estimated using the least squares method at the SNR of 5 dB; (c) linear prediction vector estimated with the ℓ_2/ℓ_1 optimization at the SNR of 5 dB ($\delta=0.1$).

By means of an auxiliary vector $\mathbf{u}$, Eq. (7) can be equivalently formulated as

$$\min_{\mathbf{a},\mathbf{u}}\left\{\frac{1}{2}\|\mathbf{X}(n)\mathbf{a}-\mathbf{x}(n)\|_{\ell_2}^2+\lambda\|\mathbf{u}\|_{\ell_1}:\mathbf{a}-\mathbf{u}=\mathbf{0}\right\}, \tag{8}$$

which has an augmented Lagrangian subproblem formulated as

$$\min_{\mathbf{a},\mathbf{u}}\left\{\frac{1}{2}\|\mathbf{X}(n)\mathbf{a}-\mathbf{x}(n)\|_{\ell_2}^2+\lambda\|\mathbf{u}\|_{\ell_1}+\boldsymbol{\eta}^T(\mathbf{a}-\mathbf{u})+\frac{\beta}{2}\|\mathbf{a}-\mathbf{u}\|_{\ell_2}^2\right\}, \tag{9}$$

where $\boldsymbol{\eta}$ is a Lagrangian multiplier vector and $\beta>0$ is a penalty parameter. The augmented term, i.e., the fourth term within the braces of Eq. (9), is introduced to ensure that the objective function is strictly convex. Given $(\mathbf{u}_k, \boldsymbol{\eta}_k)$, we can obtain $(\mathbf{a}_{k+1}, \mathbf{u}_{k+1}, \boldsymbol{\eta}_{k+1})$ by alternating minimization of Eq. (9) with respect to one variable while keeping the other variables fixed. First, for $\mathbf{u}=\mathbf{u}_k$ and $\boldsymbol{\eta}=\boldsymbol{\eta}_k$, the minimization of Eq. (9) with respect to $\mathbf{a}$ is equivalent to

$$\min_{\mathbf{a}}\left\{\frac{1}{2}\|\mathbf{X}(n)\mathbf{a}-\mathbf{x}(n)\|_{\ell_2}^2+\frac{\beta}{2}\|\mathbf{a}-\mathbf{u}_k+\boldsymbol{\eta}_k/\beta\|_{\ell_2}^2\right\}, \tag{10}$$

which has the following solution:

$$\mathbf{a}_{k+1}=[\mathbf{X}^T(n)\mathbf{X}(n)+\beta\mathbf{I}]^{-1}[\mathbf{X}^T(n)\mathbf{x}(n)+\beta\mathbf{u}_k-\boldsymbol{\eta}_k]. \tag{11}$$

Second, when $\mathbf{a}=\mathbf{a}_{k+1}$ and $\boldsymbol{\eta}=\boldsymbol{\eta}_k$ are fixed, the minimization of Eq. (9) with respect to $\mathbf{u}$ is equivalent to

$$\min_{\mathbf{u}}\left\{\lambda\|\mathbf{u}\|_{\ell_1}+\frac{\beta}{2}\|\mathbf{a}_{k+1}-\mathbf{u}+\boldsymbol{\eta}_k/\beta\|_{\ell_2}^2\right\}, \tag{12}$$

whose solution can be formulated by a soft-thresholding operator, i.e.,

$$\mathbf{u}_{k+1}=\text{soft}(\mathbf{a}_{k+1}+\boldsymbol{\eta}_k/\beta,\lambda/\beta), \tag{13}$$

where the soft function is defined as

$$\text{soft}(\boldsymbol{\xi},\mu)=\text{sgn}(\boldsymbol{\xi})\odot\max\{|\boldsymbol{\xi}|-\mu,0\},\quad\forall\boldsymbol{\xi}\in\mathbb{R}^K,\quad\mu>0. \tag{14}$$

$\text{sgn}(\cdot)$ is the signum function, $\odot$ denotes the dot product of two vectors, all the other operations are performed in a

component-wise way, and $0 \times (0/0) = 0$ is assumed. Finally, the Lagrangian multiplier vector $\boldsymbol{\eta}$ is updated by

$$\boldsymbol{\eta}_{k+1} = \boldsymbol{\eta}_k + \beta(\mathbf{a}_{k+1} - \mathbf{u}_{k+1}). \tag{15}$$

In summary, the solution of Eq. (8) can be achieved by iteratively calculating Eqs. (11), (13), and (15). Then, we obtain the prewhitened version of a microphone signal via Eq. (2). It is found from Fig. 1(c) that we can obtain a sparse prediction vector via this ℓ_2/ℓ_1-norm optimization algorithm in noisy environments.

Once the microphone signals are preprocessed by the ℓ_2/ℓ_1-norm based linear prediction, TDE can be achieved by finding the maximum of the CC function between the prediction error signals.

B. Unification of PHAT and CC from a sparse linear prediction perspective

It is found from Eq. (7) that the parameter λ plays an important role in controlling the sparseness level of the prediction vector. This parameter is mostly affected by the microphone signals, i.e., $\mathbf{X}(n)$ and $\mathbf{x}(n)$. So, we can determine λ by the following choice:

$$\lambda = \delta \|\mathbf{X}^T(n)\mathbf{x}(n)\|_{\ell_\infty}, \tag{16}$$

where $||\mathbf{z}||_{\ell_\infty} = \max_i |z_i|$ denotes the ℓ_∞-norm for any vector $\mathbf{z}$ and δ is a positive number.

(1) $\delta \to 0$: in this case, the ℓ_2/ℓ_1 optimization problem is degenerated to the traditional least squares one. So, the microphone signals are prewhitened by the least squares criterion. This prewhitening can obtain a similar effect of the preprocessing in the GCC-PHAT algorithm on the microphone signals. Figures 2(b) and 2(c) illustrate the prediction error signals via the least squares- and ℓ_2/ℓ_1-norm-based linear prediction with $\delta = 0.01$. It is found that the two prediction algorithms achieve similar whitening effect.

(2) $\delta \to \infty$: in this situation, the optimal solution of Eq. (8) tends to zero. In this case, the prediction error signal is the same as the microphone signal $\mathbf{x}(n)$. Figure 2(d) depicts the prediction error signal via the ℓ_2/ℓ_1-norm-based linear prediction with $\delta = 1.0$. We can see that the prediction error signal is comparable to the microphone signal even at $\delta = 1.0$.

(3) If δ takes some moderate value, the microphone signals are partially whitened via the ℓ_2/ℓ_1- norm based linear prediction.

Therefore, we can see that the sparse linear prediction based TDE algorithm can obtain a tradeoff between the CC and PHAT algorithms, and the PHAT and CC algorithms are unified in the same framework from the sparse linear prediction perspective.

III. SIMULATIONS

In this section, we investigate the performance of the proposed ℓ_2/ℓ_1-norm based linear prediction (ℓ_2/ℓ_1-LP) TDE algorithm. We also compare the TDE performance of the proposed algorithm, the least squares based linear prediction (LS-LP) algorithm, the CC algorithm, and the popular GCC-PHAT algorithm.[1,8] For the first two algorithms, the predictor length K is set to 128. For the proposed algorithm, the initial values of vector $\mathbf{u}$ and $\boldsymbol{\eta}$ are a null vector, respectively, the iteration times is set to 50, and $\beta = 1.0$.

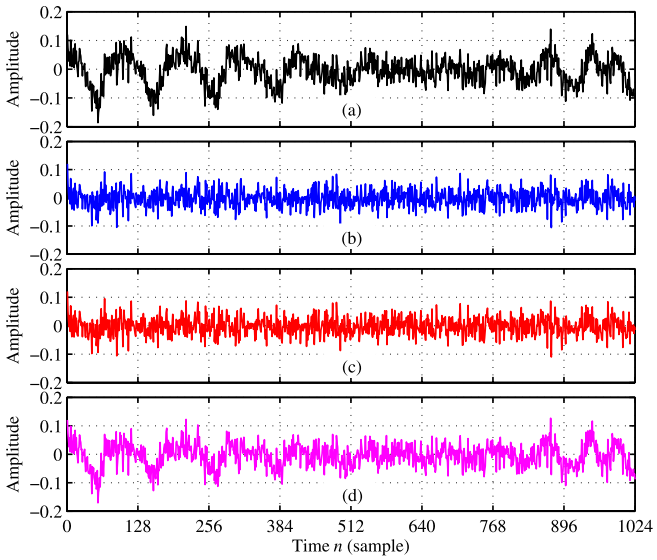

FIG. 2. (Color online) The effect of sparse penalty with different levels on the prediction error. (a) The noisy speech signal captured by a microphone; (b) the prediction error signal after the noisy speech signal is preprocessed by the least squares criterion; (c) the prediction error signal after the noisy speech signal is preprocessed by the ℓ_2/ℓ_1-norm criterion ($\delta = 0.01$); (d) the prediction error signal after the noisy speech signal is preprocessed by the ℓ_2/ℓ_1-norm criterion ($\delta = 1.0$).

Experiments are carried out in a simulated room of size 7 m × 6 m × 3 m. For ease of exposition, positions in the room are designated by (x, y, z) coordinates with reference to the southwest corner of the room floor. Two microphones are placed at (3.45, 3.00, 1.40) and (3.55, 3.00, 1.40), respectively. The sound source is located at (3.83,1.50,1.40). The impulse responses from the source to the two microphones are generated using the image model.[19] The microphones' outputs are obtained by convolving the source signal with the corresponding generated impulse responses and then adding zero-mean white Gaussian noise to the results to control the signal-to-noise ratio (SNR).

In the simulations, the microphone signals are partitioned into nonoverlapping frames with frame length of 64 ms. We use the probability of anomalous estimates and the root mean square error (RMSE) of nonanomalous estimates[2,3] to evaluate the performance of the proposed algorithm. The source signal is a segment of speech signal from a male talker, which is sampled at 16 kHz, and the length of the signal is approximately 1 min. The total number of frames is 950 (the frame length is 1024 samples). The true time delay from the sound source to the two microphones is −1.0 samples.

Figure 3 plots the TDE results versus SNR in reverberant environments (the reverberation time $T_{60} = 120, 300$ ms). It is seen from Figs. 3(a) and 3(b) that the LS-LP TDE algorithm is comparable or slightly superior to the PHAT

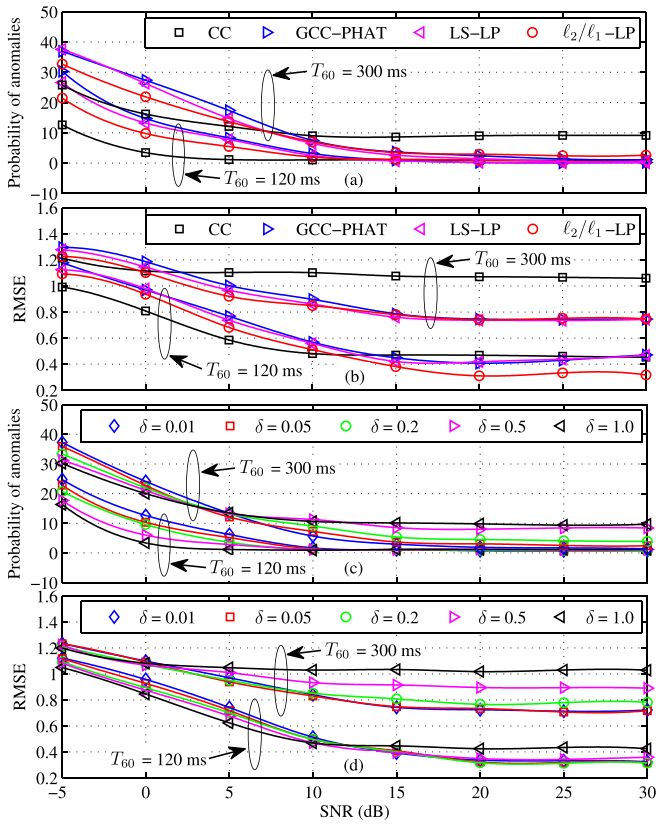

FIG. 3. (Color online) The probability of anomalous time delay estimates and RMSE of nonanomalous time delay estimates in noisy and reverberant environments, where (a) and (b) show the TDE results of the four algorithms (for the ℓ_2/ℓ_1-LP algorithm, $\delta = 0.1$); (c) and (d) illustrate the TDE results of the proposed ℓ_2/ℓ_1-LP algorithm with different values of δ.

algorithm, which indicates that the linear prediction is effective to whiten the microphone signals for TDE. As far as the three prewhitening based TDE algorithms are concerned, the ℓ_2/ℓ_1-LP algorithm ($\delta = 0.1$) obtains the best TDE performance under the noisy and lightly or moderately reverberant environments. The proposed ℓ_2/ℓ_1-LP algorithm displays its robustness to noise at low SNRs due to introduction of the sparse constraint to the prediction vector. When SNR is low, the ℓ_2/ℓ_1-LP algorithm obtains a compromise beween the CC and prewhitening-based algorithms.

Figures 3(c) and 3(d) plot the TDE results of the proposed algorithm with different values of δ in noisy and reverberant environments ($T_{60} = 120$, 300 ms). It is seen that as δ increases, the prediction vector becomes more sparse, and so the ℓ_2/ℓ_1-LP algorithm is more robust to noise, while more and more sensitive to reverberation. Thus, a proper value of δ needs to be found in practical applications, depending on the level of noise and reverberation.

IV. CONCLUSIONS

In this paper, we developed an ℓ_2/ℓ_1-norm based linear prediction optimization algorithm for TDE. This algorithm unifies the cross correlation method without prewhitening and that with prewhitening and it also provides a tradeoff between the traditional CC and PHAT algorithms in dealing with noise and reverberation. Experiments were carried out and the results showed that the TDE algorithm with light prewhitening of the microphone signals is robust to noise, while the TDE algorithm with heavy prewhitening is robust to reverberation.

ACKNOWLEDGMENTS

H.H. was supported by the Open Foundation of the Key Laboratory of Modern Acoustics of Nanjing University (Grant No. 1302), the Incubation Program for the Distinguished Youth Foundation of Sichuan Province of China (Grant No. 2014JQ0042), and the Doctoral Foundation of Southwest University of Science and Technology (Grant No. 13zx7149). T.Y. was partially supported by the Open Foundation of Robot Technology Used for Special Environment Key Laboratory of Sichuan Province (Grant No. 13zxtk06).

[1]C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," IEEE Trans. Acoust. Speech Sign. Process. **ASSP-24**, 320–327 (1976).
[2]J. Chen, J. Benesty, and Y. Huang, "Robust time delay estimation exploiting redundancy among multiple microphones," IEEE Trans. Speech Audio Process. **11**, 549–557 (2003).
[3]H. He, L. Wu, J. Lu, X. Qiu, and J. Chen, "Time difference of arrival estimation exploiting multichannel spatio-temporal prediction," IEEE Trans. Audio Speech Lang. Process. **21**, 463–475 (2013).
[4]J. Chen, J. Benesty, and Y. Huang, "Time delay estimation in room acoustic environments: An overview," EURASIP J. Appl. Sign. Process. **2006**, 1–19 (2006).
[5]S. Doclo and M. Moonen, "Robust adaptive time delay estimation for speaker localization in noisy and reverberant acoustic environments," EURASIP J. Appl. Sign. Process. **2003**, 1110–1124 (2003).
[6]J. Benesty, Y. Huang, and J. Chen, "Time delay estimation via minimum entropy," IEEE Sign. Process. Lett. **14**, 157–160 (2007).
[7]H. He, J. Lu, L. Wu, and X. Qiu, "Time delay estimation via non-mutual information among multiple microphones," Elsevier Appl. Acoust. **74**, 1033–1036 (2013).
[8]Y. Huang, J. Benesty, and J. Chen, *Acoustic MIMO Signal Processing* (Springer, Berlin, 2006), Chap. 9, pp. 215–259.
[9]J. Makhoul, "Linear prediction: A tutorial review," Proc. IEEE **63**, 561–580 (1975).
[10]B. Yegnanarayana, S. R. M. Prasanna, R. Duraiswami, and D. Zotkin, "Processing of reverberent speech for time-delay estimation," IEEE Trans. Speech Audio Process. **13**, 1110–1118 (2005).
[11]E. D. Claudio and R. Parisi, "Multi-source localization strategies," in *Microphone Arrays: Signal Processing Techniques and Applications*, edited by M. Brandstein and D. Ward (Springer, New York, 2001), pp. 181–201.
[12]R. P. Ramachandran and P. Kabal, "Pitch prediction filters in speech coding," IEEE Trans. Acoust. Speech Sign. Process. **ASSP-37**, 467–478 (1989).
[13]D. Giacobello, M. G. Christensen, M. N. Murthi, S. H. Jensen, and M. Moonen, "Sparse linear prediction and its applications to speech processing," IEEE Trans. Audio Speech Lang. Process. **20**, 1644–1657 (2012).
[14]J. F. Yang and Y. Zhang, "Alternating direction algorithms for l_1-problems in compressive sensing," SIAM J. Sci. Comput. **33**, 250–278 (2011).
[15]W. Cheney and D. Kincaid, *Numerical Mathematics and Computing*, 6th ed. (Brooks/Cole Thomson Learning, Belmont, CA, 2008), pp. 1–723.
[16]X. Jiang, T. Kirubarajan, and W. J. Zeng, "Robust sparse channel estimation and equalization in impulsive noise using linear programming," Elsevier Sign. Process. **93**, 1095–1105 (2013).
[17]S. J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky, "An interior-point method for largescale l_1-regularized least squares," IEEE J. Select. Top. Sign. Process. **1**, 606–617 (2007).
[18]S. J. Wright, *Primal-Dual Interior-Point Methods* (SIAM, Philadelphia, PA, 1997), pp. 1–289.
[19]J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," J. Acoust. Soc. Am. **65**, 943–950 (1979).