BLIND SPARSE-NONNEGATIVE (BSN) CHANNEL IDENTIFICATION FOR ACOUSTIC TIME-DIFFERENCE-OF-ARRIVAL ESTIMATION

Yuanqing Lin[†], *Jingdong Chen*[‡], *Youngmoo Kim*[♯], *and Daniel D. Lee*[†]

[†]GRASP Laboratory, Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, PA 19104

[‡]Bell Laboratories, Alcatel-Lucent, 600 Mountain Avenue, Murray Hill, NJ 07974

[♯]Electrical and Computer Engineering Department, Drexel University, Philadelphia, PA 19104

linyuanq, ddlee@seas.upenn.edu, jingdong@research.bell-labs.com, ykim@drexel.edu

ABSTRACT

Estimating time-difference-of-arrival (TDOA) remains a challenging task when acoustic environments are *reverberant* and *noisy*. Blind channel identification approaches for TDOA estimation explicitly model multipath reflections and have been demonstrated to be effective in dealing with reverberation. Unfortunately, existing blind channel identification algorithms are sensitive to ambient noise. This paper shows how to resolve the noise sensitivity issue by exploiting prior knowledge about an acoustic room impulse response (RIR), namely, an acoustic RIR can be modeled by a *sparse-nonnegative* FIR filter. This paper shows how to formulate a single-input two-output blind channel identification into a least square *convex* optimization, and how to incorporate the sparsity and nonnegativity priors so that the resulting optimization remains convex and can be solved efficiently. The proposed *blind sparse-nonnegative (BSN) channel identification* approach for TDOA estimation is not only robust to reverberation, but also robust to ambient noise, as demonstrated by simulations and experiments in real acoustic environments.

1. INTRODUCTION

Time delay estimation [1], which calculates the time-difference-of-arrival (TDOA) between signals received at different microphones, is essential for sound source localization using microphone arrays. The task of TDOA estimation is illustrated in Fig. 1. In terms of the underlying model for an acoustic room impulse response (RIR), the existing approaches for TDOA estimation can be classified into two categories: generalized cross-correlation (GCC) approaches and blind channel identification approaches. The GCC approaches approximate an acoustic RIR as a simple delta function, and the TDOA estimation is achieved by maximizing some weighted cross-correlation function with respect to a scalar time difference. An excellent review of this category of approaches can be found in [2]. The GCC approaches do not explicitly take multipath reflections into account and their performance in reverberant acoustic environments is limited due to the underlying unrealistic RIR model. In contrast, blind channel identification approaches [3] [4] model an acoustic RIR as an FIR filter that includes both a direct path and multipath reflections. In these approaches, after the modeling filters have been identified, the TDOA can be easily computed by examining the direct paths in the filters. By using a more realistic model, the blind channel identification approaches have been shown to be more effective than GCC approaches to reverberation. Unfortunately, blind channel identification approaches have been found to be sensitive to ambient noise.

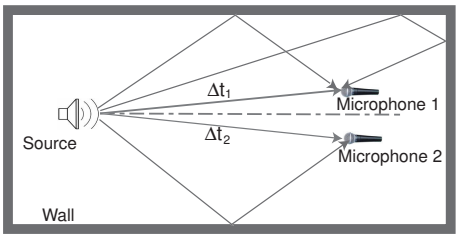


Figure 1: Illustration of a single-input two-output acoustic system. A microphone observation consists of a direct path signal, multipath reflections, and ambient noise. The task of TDOA estimation is to estimate the time difference of arrival between the two direct paths, $\Delta t_2 - \Delta t_1$.

This is because blind channel identification needs to estimate a much more complex model having hundreds or even thousands of parameters (filter coefficients) and is often ill-conditioned due to the nature of blind estimation.

This paper proposes to resolve the noise sensitivity issue in blind channel identification by exploiting prior knowledge about acoustic RIRs. According to many studies [5], an acoustic RIR can be modeled by an FIR filter, which is both *nonnegative* and *sparse* in theory. In practice, nonnegativity and sparsity may not be strictly satisfied due to effects such as low- or high-pass filtering in the propagation media or the imperfect frequency response of a microphone. However, when those effects are common to both channels, they can be viewed as distortions to a common source. Therefore, the nonnegativity and sparsity assumption are reasonable for real acoustic environments if an acoustic system is appropriately constructed.

The nonnegativity and sparsity priors have been demonstrated to be effective in many signal processing tasks [6]. Our previous work [7] showed that these two priors provided dramatic regularization to the least-mean-square (LMS) problem for identifying acoustic RIRs and improved its robustness to ambient noise when the source was given *a priori*. This paper shows that they play a critical role in *blind* acoustic channel identification for resolving ill-conditioned solutions, which may be caused by overestimating the filter length or insufficient excitation due to the band-limited nature of speech sources [8]. By making the problem better posed, the resulting blind sparse-nonnegative (BSN) channel identification approach is robust to ambient noise. Furthermore, the BSN channel identification approach also allows common preprocessing on the microphone observations to reduce the noise level. In

contrast, conventional blind channel identification approaches prohibit preprocessing since they are not able to resolve the preprocessing filtering from filtering by a RIR.

2. BLIND SPARSE-NONNEGATIVE (BSN) CHANNEL IDENTIFICATION

2.1. Previous work

In an acoustic system as illustrated in Fig. 1, the microphone outputs at time k can be written as:

$$x_i(k) = s(k) * h_i + n_i(k), \;\; i = 1, 2, \tag{1}$$

where $*$ denotes linear convolution, $s(k)$ is a source signal, h_i represents the channel impulse response between the source and the ith microphone, and $n_i(k)$ is a noise signal. The blind channel identification via *cross relation* is based on a clever observation, $x_2(k) * h_1 = x_1(k) * h_2 = s(k) * h_1 * h_2$, if the microphone signals are noiseless [8]. Then, without requiring any knowledge from the source signal, the channel filters can be identified by minimizing the squared cross relation error. In matrix-vector form, the optimization becomes

$$\begin{aligned} \mathbf{h}_1^*, \mathbf{h}_2^* &= \arg\min_{\mathbf{h}_1, \mathbf{h}_2} \frac{1}{2} \|\mathbf{X}_2\mathbf{h}_1 - \mathbf{X}_1\mathbf{h}_2\|^2 \\ &\quad \text{subject to } \; \|\mathbf{h}_1\|^2 + \|\mathbf{h}_2\|^2 = 1, \end{aligned} \tag{2}$$

where $\mathbf{X}_i$ is the $(N + L - 1) \times L$ convolution Toeplitz matrix whose first row and first column are $[x_i(k - N + 1), x_i(k - N), \ldots, x_i(k - N - L + 2)]$ and $[x_i(k - N + 1), x_i(k - N + 2), ..., x_i(k), 0, \ldots, 0]^T$, respectively, N is the microphone signal length, L is the filter length, $\| \cdot \|$ denotes l_2-norm, and the constraint is to avoid the trivial zero solution. It is easy to see that the above optimization is a minimum eigenvalue problem, and it can be solved by eigenvalue decomposition. Benesty [3] proposed to solve the above optimization in an adaptive way, and demonstrated that the algorithm is effective in dealing with reverberation for TDOA estimation. Unfortunately, the filter estimation by the optimization problem in Eq. 2 is sensitive to ambient noise.

To improve the robustness to ambient noise, our strategy is to incorporate blind channel identification with prior knowledge about an acoustic RIR, namely, an acoustic RIR can be modeled by a *sparse-nonnegative* FIR filter. However, it is hard to incorporate either the nonnegativity prior or the sparsity prior directly into the optimization in Eq. 2. In fact, if the optimization in Eq. 2 is also subject to nonnegative constraints, the resulting optimization is NP-hard. Consequently, we choose to reformulate the blind channel identification into a *convex* optimization problem, which will provide a flexible platform for incorporating both the nonnegativity prior and the sparsity prior. We will focus on the batch-mode formulation in this paper and show its adaptive counterpart in future work.

2.2. Convex formulation

The optimization in Eq. 2 is not convex because its domain, $\|\mathbf{h}_1\|^2 + \|\mathbf{h}_2\|^2 = 1$, is not convex [9]. However, this non-convex constraint which is used to avoid the trivial zero solution, can be replaced by a convex constraint, which is also able to avoid the trivial zero solution. Our choice is a singleton linear constraint and the optimization becomes

$$\begin{aligned} \mathbf{h}_1^*, \mathbf{h}_2^* &= \arg\min_{\mathbf{h}_1, \mathbf{h}_2} \frac{1}{2} \|\mathbf{X}_2\mathbf{h}_1 - \mathbf{X}_1\mathbf{h}_2\|^2 \\ &\quad \text{subject to } \; h_1(0) = 1, \end{aligned} \tag{3}$$

where $h_1(0)$ is the first element of $\mathbf{h}_1$. Because the optimization is a minimization, its solution tends to align $h_1(0)$ with the maximum coefficient in the filter $\mathbf{h}_1$, which is often the coefficient corresponding to the direct path. To ensure that $\mathbf{h}_2$ does not have nonzero elements at negative time delays, one can use earlier samples for x_1. How much earlier is determined by the maximum possible time delay, $f_s \cdot d/c$, where d is the distance between the two microphones, f_s is sampling rate, and c is the speed of sound in air. It can be shown that, when the microphone signals are noiseless, the two optimizations (Eqs. 2 and 3) yield equivalent solutions up to a constant delay and a constant scalar factor.

The new formulation in Eq. 3 has many advantages. It is convex, and can be written as an unconstrained least square (LS) problem since the singleton constraint can be easily substituted into the objective function. Furthermore, the resulting LS approach is more robust to ambient noise than the eigenvalue decomposition approach in Eq. 2. This can be better seen in the frequency domain. The squared cross relation error is weighted by the power spectrum density of the underlying common source. As a result, when microphone signals are noisy, the optimization in Eq. 2 tends to fill the filter energy constraint with less significant frequency bands which have little contribution in the source. This is because the squared error in those frequency bands are weighted less in the objective function. Consequently, the solution to Eq. 2 is extremely sensitive to ambient noise. In contrast, the singleton linear constraint in Eq. 3 has much less coupling in the filter energy allocation, and thus its solution is more robust to ambient noise.

2.3. BSN channel identification algorithm

The convex LS formulation in Eq. 3 provides a flexible platform for incorporating the nonnegativity and sparsity priors. The optimization for *blind sparse-nonnegative (BSN) channel identification* becomes

$$\begin{aligned} \mathbf{h}_1^*, \mathbf{h}_2^* &= \arg\min_{\mathbf{h}_1, \mathbf{h}_2} \frac{1}{2} \|\mathbf{X}_2\mathbf{h}_1 - \mathbf{X}_1\mathbf{h}_2\|^2 + \lambda' \sum_{j=0}^{L-1} [h_1(j) + h_2(j)] \\ &\quad \text{subject to } \; h_1(0) = 1, \mathbf{h}_1 \geq 0, \mathbf{h}_2 \geq 0 \end{aligned} \tag{4}$$

where the second term is the l_1-norm of the filters, and λ' is the sparsity regularization parameter that balances the preference between the squared fitting error and the sparseness of the solution described by its l_1-norm. Enforcing sparsity using l_1-norm regularization has been an active research area in the last decade [6], and it has been the driving force for many emerging fields in signal processing, such as sparse coding and compressive sensing. As for the nonnegative constraints, they were inspired by nonnegative matrix factorization (NMF) [10], which showed that nonnegative constraints are able to dramatically regularize an optimization problem. Combining both the nonnegative constraints and the l_1-norm regularization, the optimization in (4) is expected to resolve the ill-conditioning problem in blind channel identification and yield solutions that are robust to ambient noise.

Given a sparsity regularization parameter λ', the optimization in Eq. 4 is a convex nonnegative quadratic programming (NNQP)

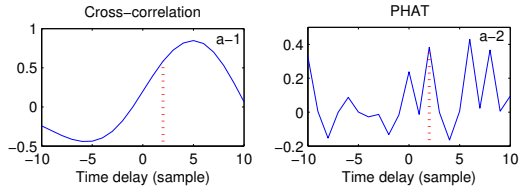
problem, which can be solved by various methods with guaranteed global convergence. Among those, the multiplicative update algorithm [11] is able to solve the NNQP problem efficiently since it only involves Toeplitz matrix-vector multiplication, which can be implemented by FFTs. Another important issue in Eq. 4 is how to determine the regularization parameter λ', which controls the sparseness of solutions. The work in [7] shows that, in the Bayesian framework, the optimal regularization parameter λ' is equal to the product $\sigma^2\lambda$, where σ^2 describes the noise level and λ is the parameter describes the sparseness of filters. These two parameters can be determined by either *a priori* knowledge, or learning from observed microphone signals [7].

3. RESULTS

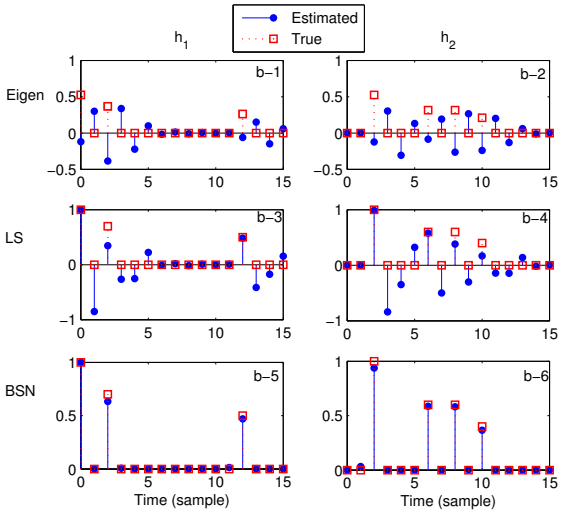
3.1. A simulated example

Here we first provide a toy example to illustrate the the advantage of the proposed BSN channel identification approach for TDOA estimation in comparison with other existing approaches. In the simulation, the source (s) is a speech segment of 4096 samples with sampling rate of 16 kHz, and both of the two FIR filters (h_1 and h_2) are 16 samples long. If we use $j = 0, 1, ...15$ to index the filter coefficients, filter h_1 has nonzero elements only at j=0, 2, and 12 with amplitudes of 1, 0.7, and 0.5, respectively; filter h_2 has nonzero elements only at j=2, 6, 8, and 10 with amplitudes of 1, 0.6, 0.6 and 0.4, respectively. Notice that both filters are nonnegative and sparse. Then, the simulated microphone observations (x_i) were computed according to Eq. 1 where the ambient noise (n_i) was real noise recorded in a conference room. The noise was scaled so that the signal-to-noise ratio (SNR) of the microphone signals was 15 dB. The simulated microphone signals were then highpassed with a cut-off frequency of 300 Hz to reduce the low frequency noise before they were fed to different algorithms for TDOA estimation.

The simulation results are shown in Fig. 2. The traditional cross-correlation approach [Fig. 2 (a-1)] has low temporal resolution, and multipath reflections often cause a peak shift in the cross-correlation function. Consequently, this approach performs poorly in reverberant environments. The phase transform (PHAT) approach [Fig. 2 (a-2)] improves the temporal resolution by prewhitening the microphone signals, however, its performance is still limited by the underlying oversimplified RIR model. The simulation results of blind channel identification approaches are shown in Fig. 2 (b), illustrating strong advantages of our new formulation of blind channel identification presented in Section 2. As shown in Fig. 2 (b), the new LS formulation in Eq. 3 is more robust to ambient noise than the conventional eigenvalue decomposition approach in Eq. 2. Moveover, the sparsity and nonnegativity prior knowledge helps to resolve the degeneracy in blind channel identification and yields dramatic improvement in filter estimates. The filter estimation accuracy gained by the BSN channel identification approach will become critical when the filters are thousands of taps long, as in typical real acoustic environments.



(a) GCC approaches. In each figure, the solid line describes the GCC function between two microphone signals, and the vertical dot line indicates the true time delay. The traditional cross-correlation is on the left and the phase transform (PHAT) is on the right.



(b) Blind channel identification approaches. The three rows from top to bottom are the identified filters respectively by eigenvalue decomposition approach (Eq. 2), LS approach (Eq. 3) and the BSN channel identification approach (Eq. 4). The left and right columns represent the identified filters associated with channel 1 and channel 2, respectively. In each figure, the dot-solid line describes the identified filters, and the square-dot line indicates the true filters up to a constant time delay and a constant scalar factor.

Figure 2: Results of GCC approaches and blind channel identification approaches for TDOA estimation.

3.2. Performance comparison using real room recordings

Now we evaluate the performance of the proposed BSN channel identification approach for TDOA estimation in real environments. The experimental setup is illustrated in Fig. 3. Prerecorded speech sequences were played through a loudspeaker located at one end of the room and recorded by a matched omnidirectional microphone pair (SP-CMC-8, Sound Professionals) located at the other end of the room. We recorded two data sets: one set had the loudspeaker in the middle (see Position 1 in Fig. 3), and the other had the loudspeaker about 75 cm away from the middle (see Position 2 in Fig. 3). At each speaker position, 100 speech sentences (50 by a male speaker and 50 by a female speaker) were played and recorded with a sampling rate of 16 kHz. In our evaluation, we divided the recordings into segments of 4096 samples, and discarded those silent segments which contained no speech signals. Then, we treated each segment independently and performed TDOA algorithms on each of them. Since a large portion of the ambient noise was at low frequency (such as air-conditioning noise), the recorded signals were highpassed with a cut-off frequency of 300 Hz before they were fed to TDOA estimation algorithms. For the BSN channel identification approach, the filter length was 2048.

As shown in Fig. 4, the proposed BSN channel identification approach yielded consistent TDOA estimates at both Position 1 and Position 2, even though Position 2 is difficult for TDOA es-

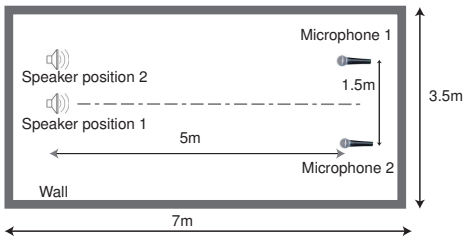


Figure 3: The loudspeaker-microphone positions in a conference room during recording. The dot-dash line indicates the center line of the room.

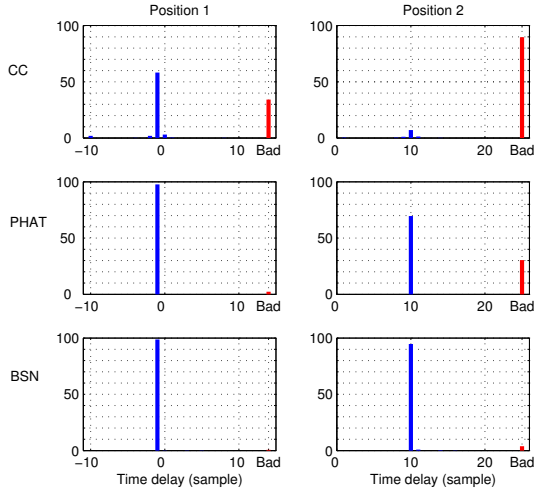


Figure 4: Histogram in percentage of TDOA estimates using three different approaches: the cross-correlation (CC) approach, the phase transform (PHAT) approach, and the BSN channel identification approach. The left and right column describes the TDOA estimation results when the speaker was at Position 1 and Position 2, respectively. The bad estimates are those that are more than 10 samples away from the true values (-1 for Position 1, and 10 for Position 2).

timation since the loudspeaker was close to the wall and the wall reflections were very strong. In contrast, the PHAT approach had good estimates only at position 1 but not position 2. The cross-correlation approach did not yield satisfactory estimates at either positions and almost completely failed at position 2. As for other blind channel identification approaches, the batch-mode eigenvalue decomposition (in Eq. 2) and the LS (in Eq. 3), they were not able to yield competitive results simply because there were not enough frequency components in a short 4096-sample frame for estimating filters of length 2048. The BSN channel identification approach overcomes the difficulty by exploiting knowledge about the nonnegativity and sparsity of the RIRs.

4. DISCUSSION

We have developed a blind sparse-nonnegative (BSN) channel identification approach for TDOA estimation, which exploits prior knowledge about an acoustic RIR, namely, an acoustic RIR can be modeled by a sparse-nonnegative FIR filter. The BSN channel identification is formulated as an l_1-norm regularized nonnegative LS problem, which is convex and can be solved efficiently with guaranteed global convergence. Both simulation and experimental results in real acoustic environments demonstrate the effectiveness of the BSN channel identification approach for TDOA estimation.

Although modeling an acoustic RIR as a sparse-nonnegative FIR filter is demonstrated to be effective for TDOA estimation, how accurate the modeling is in real acoustic environments remains an open problem. TDOA estimation is relatively immune to moderate modeling inaccuracy since it only requires information about the direct path but not the whole filter. Nevertheless, we believe exploiting prior knowledge about RIRs is crucial for blind channel identification to resolve its underlying degeneracy and become robust to ambient noise.

Our future work is to develop an adaptive algorithm for BSN channel identification. We expect the resulting adaptive algorithm would outperform the adaptive eigenvalue decomposition (AED) algorithm [3], which has been shown to be not only computationally efficient, but also effective in dealing with reverberation.

5. REFERENCES

[1] J. Chen, J. Benesty, and Y. A. Huang, "Time delay estimation in room acoustic environments: An overview," *EURASIP Journal on Applied Signal Processing*, vol. 2006, pp. Article ID 26 503, 19 pages, 2006.

[2] C. H. Knapp and G. C. Carter., "The generalized correlation method for estimation of time delay," *IEEE Trans. ASSP*, vol. 24, no. 4, pp. 320–327, 1976.

[3] J. Benesty, "Adaptive eigenvalue decomposition algorithm for passive acoustic source localization," *J. Acoust. Soc. Am.*, vol. 107, no. 1, pp. 384–391, 2000.

[4] S. Doclo and M. Moonen, "Robust adaptive time delay estimation for speaker localization in noisy and reverberant acoustic environments," *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 11, pp. 1110–1124, 2003.

[5] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.*, vol. 65, pp. 943–950, 1979.

[6] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Scientific Computing*, vol. 20, no. 1, pp. 33–61, 1998.

[7] Y. Lin and D. D. Lee, "Bayesian Regularization And Nonnegative Deconvolution (BRAND) for room impulse response estimation," *IEEE Trans. Signal Processing*, vol. 54, no. 3, pp. 839–847, 2006.

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

[9] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[10] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[11] F. Sha, Y. Lin, L. K. Saul, and D. D. Lee, "Multiplicative updates for nonnegative quadratic programming," *Neural Comput.*, vol. 19, no. 8, pp. 2004–2031, 2007.