

Employing Sparsity for Joint Sound Source & Acoustic Channel Estimation

Youngmoo E. Kim and Travis M. Doll
Electrical and Computer Engineering · Drexel University

1 Abstract

All audio is experienced through an acoustic channel, which imposes some amount of alteration upon the “true” source signal. Most often, this is due to the characteristics of the environment (e.g., the physical configuration of a room or space). Assuming this to be a linear process, the sound we hear is a convolution of the “true” source and the impulse response of the acoustic channel. In prior work [1, 2] we have investigated the use of sparse methods for blind room channel (impulse response) estimation, which tend to be sparse in the time domain [3]. Currently, we are pursuing an extension of this work to incorporate a sparse model for describing the acoustic source as well, particularly for musical sources, and to estimate these source and channel model parameters jointly. In doing so, we hope to obtain an improved estimate of the underlying source signal simultaneously with the impulse response of the channel. This method has applications for musical instrument recognition and coding, and an accurate estimate of the acoustic channel could be potentially beneficial for solving problems of music transcription from live (reverberant) sound data.

2 Previous framework

Our previous efforts have focused on blind channel estimation, in which no prior information is known regarding the source or the channel. By itself, the problem is degenerate and can not be solved. A possible solution stems from having two observations (microphones) of the same source from different locations in a room (and thus using two different acoustic channels) and assuming the sparsity of the room impulse responses. If, for example, we have two observations $x_1[n]$ and $x_2[n]$:

$$x_i[n] = s[n] * h_i[n], \quad i = 1, 2, \quad (1)$$

where $s[n]$ is the the “true” source signal and $h_i[n]$ is the acoustic channel for the i th observation. We can also express this in linear algebra form:

$$\mathbf{x}_i = \mathbf{H}_i \cdot \mathbf{s} \quad (2)$$

where \mathbf{H}_i is a convolution Toeplitz matrix whose columns are row-shifted values of $h_i[n]$ and \mathbf{s} is a vector containing samples of $s[n]$. Having two observations, we also know that

$$x_2[n] * h_1[n] = s[n] * h_2[n] * h_1[n] = x_1[n] * h_2[n], \quad (3)$$

meaning that the convolutions of each observation with the acoustic channel from the other observation should be equivalent. Again, this can be expressed in matrix form as

$$\mathbf{X}_2(\mathbf{s}) \cdot \mathbf{h}_1 = \mathbf{X}_1(\mathbf{s}) \cdot \mathbf{h}_2. \quad (4)$$

Here, $\mathbf{X}_i(\mathbf{s})$ is a convolution Toeplitz matrix whose columns are row-shifted values of \mathbf{x}_i and \mathbf{h}_i are vectors containing the filter coefficients $h_i[n]$, and the relationship with the source \mathbf{s} is explicitly notated as a dependent variable. The ideal equivalence of these quantities implies that minimizing the difference between the two will yield estimates for the acoustic channels \mathbf{h}_1^* and \mathbf{h}_2^* . Therefore, the solution can be presented as the following optimization problem:

$$\mathbf{h}_1^*, \mathbf{h}_2^* = \arg \min_{\mathbf{h}_1, \mathbf{h}_2} \frac{1}{2} \|\mathbf{X}_2(\mathbf{s}) \cdot \mathbf{h}_1 - \mathbf{X}_1(\mathbf{s}) \cdot \mathbf{h}_2\|_2^2 + \lambda'(|\mathbf{h}_1| + |\mathbf{h}_2|), \quad (5)$$

$$\text{such that } h_1(0) = 1. \quad (6)$$

Although we use an l_2 -norm for the primary minimization objective, the l_1 -norm used in the regularization formula encourages the sparsity of \mathbf{h}_i . The constraint in Eq. 6 ensures that the optimization is convex, so that unique solutions are easily found, although exact scaling information in the acoustic channel coefficients may be lost. This is not usually an issue, since the relative energies between the channels will be preserved. Determination of the regularization parameter λ can be performed in several ways, including a Bayesian estimation framework described in [2].

3 Proposed work

In many situations, particularly when dealing with musical sources, \mathbf{s} can be approximated by a linear combination of some basis vectors, collected in matrix \mathbf{B} . Upon choosing an appropriate basis [4], the vector containing the weighting coefficients, which we call $\hat{\mathbf{s}}$, that represent the source signal may also be sparse:

$$\mathbf{s} \approx \mathbf{B} \cdot \hat{\mathbf{s}} = \tilde{\mathbf{s}}. \quad (7)$$

Since \mathbf{B} is chosen a priori, from a single observation we can jointly solve for $\hat{\mathbf{s}}$ and \mathbf{h}_1 by minimizing the difference between observed signal \mathbf{x}_1 and the convolution of the estimated source and channel, i.e.:

$$\hat{\mathbf{s}}^*, \mathbf{h}_1^* = \arg \min_{\hat{\mathbf{s}}, \mathbf{h}_1} \frac{1}{2} \|\mathbf{x}_1 - \tilde{\mathbf{S}}(\hat{\mathbf{s}}) \cdot \mathbf{h}_1\|_2^2 + \lambda_1 |\mathbf{h}_1| + \lambda_2 |\hat{\mathbf{s}}|, \quad (8)$$

such that $h_1(0) = 1$.

In keeping with our notation, $\tilde{\mathbf{S}}(\hat{\mathbf{s}})$ is a convolution Toeplitz matrix whose columns are row-shifted values of $\tilde{\mathbf{s}}$. Extending the variational formulation, we introduce a second regularization on the l_1 -norm of $\hat{\mathbf{s}}$ to promote the sparsity of this vector of basis weights. This optimization remains convex, although the addition of the second regularization parameter adds significant complexity. We are currently exploring other constraints that may lead to more efficient solutions to this optimization problem, and the estimation of regularization parameters λ_1 and λ_2 remains an additional topic of ongoing research. To incorporate information from multiple observations as before, we are also investigating the following formulation of the problem:

$$\hat{\mathbf{s}}^*, \mathbf{h}_1^*, \mathbf{h}_2^* = \arg \min_{\hat{\mathbf{s}}, \mathbf{h}_1, \mathbf{h}_2} \frac{1}{2} \|\mathbf{x}_1 - \tilde{\mathbf{S}}(\hat{\mathbf{s}}) \cdot \mathbf{h}_1\|_2^2 + \|\mathbf{x}_2 - \tilde{\mathbf{S}}(\hat{\mathbf{s}}) \cdot \mathbf{h}_2\|_2^2 + \lambda_1 (|\mathbf{h}_1| + |\mathbf{h}_2|) + \lambda_2 |\hat{\mathbf{s}}|, \quad (9)$$

such that $h_1(0) = 1$.

In this formulation, however, the objective function is no longer convex, but we can use blind channel estimates \mathbf{h}_1^* and \mathbf{h}_2^* as determined using Eq. (5) to initialize this joint source-channel estimation.

References

- [1] Y. Lin, J. Chen, Y. E. Kim, and D. D. Lee, "Blind sparse-nonnegative (BSN) channel identification for acoustic time-difference-of-arrival estimation," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. New Paltz, NY: IEEE, 21-24 October 2007.
- [2] —, "Blind channel identification for speech dereverberation using l_1 -norm sparse learning," in *Proc. Conference on Neural Information Processing Systems (NIPS)*. Vancouver, Canada: NIPS, 3-6 December 2007.
- [3] J. Allen and D. Berkley, "Image method for efficiently simulating small room acoustics," in *Journal of Acoustic Society of America*, April 1979, pp. 912–915.
- [4] P.-A. Manzagol, T. Bertin-Mahieux, and D. Eck, "On the use of sparse time-relative auditory codes for music," in *Proc. International Conference on Music Information Retrieval (ISMIR)*. Philadelphia, PA: ISMIR, 14-18 September 2008.