SOUND FIELD DECOMPOSITION USING SPICE DECOMPOSITION

Satoru Emura and Noboru Harada

NTT Media Intelligence Laboratories

ABSTRACT

We propose a method to estimate a reverberant sound field by compressed sensing approach without using the hyper parameter for controlling sparsity. This method first applies sparse iterative covariance-based approach (SPICE) to the power spectral density matrix of the microphone signals and obtains the estimate of sensor noise power without using a hyper parameter. From this estimate, a convex optimization problem for a sparse solution is obtained that relates the microphone signals, plane-wave expansion coefficients, and the sensor noise power. The plane-wave expansion coefficients are obtained as the solution of this convex optimization problem. The sound field can be estimated from the plane-wave expansion coefficients.

Index Terms— Sparse, group LASSO, SPICE, convex optimization, constraint

1. INTRODUCTION

Sound pressure at several points in a sound field is measured by placing microphones at these points. Sound pressure at the other points can be estimated when the sound field is correctly estimated from the multichannel microphone signals [1]. The sound field can be estimated by using relatively few microphones when a compressed sensing (CS) approach is applied to obtain a sparse solution [2]. For example, the least absolute shrinkage and selection operator (LASSO) [3] [4] is used to obtain a sparse solution from a single measurement vector of the multichannel microphone signals. This approach was extended to multiple measurement vectors in [5] and [6]. Using multiple measurement vectors is expected to make the estimation more robust. Group LASSO [7] is often used to handle the case of the multiple measurement vectors.

However, both LASSO and group LASSO require a hyper parameter for controlling the sparsity of their solutions. The hyper parameter is not known beforehand. Usually, users of LASSO and group LASSO tune the hyper parameter by checking the sparsity of the solution several times. Hence in an ordinary reverberant room with some noise, we have to tune the hyper parameter beforehand to estimate a sound field by using the CS approach.

We propose a method to estimate a reverberant sound field based on CS approach without using the hyper parameter. This method first applies sparse iterative covariance-based approach (SPICE) [8] to the power spectral density (PSD) matrix of the microphone signals and obtains the estimate of sensor noise power without using a hyper parameter. From this estimate, a convex optimization problem for a sparse solution is obtained that relates the microphone signals, plane-wave expansion coefficients, and the sensor noise power. The plane-wave expansion coefficients are obtained as the solution of this convex optimization problem. The sound field can be estimated from the plane-wave expansion coefficients. In Section 2, we review conventional methods to obtain a sparse solution. In Section 3, we discuss our proposed method. In Section 4, we discuss the evaluation of the proposed method.

2. LASSO, GROUP LASSO, AND SPICE

We first explain LASSO and group LASSO which conventional CS-based sound-field-estimation methods use. Next, we explain SPICE on which our proposed method is based.

We formulate the problem of sound field estimation in the short-time Fourier transform (STFT) domain. Let m be the frame number. The nth microphone signal $Y_n(m,\omega)$ $(1 \leq n \leq N)$ in the frequency domain is expressed as

$$Y_n(m,\omega) = G_n(\omega)S(m,\omega) + V_n(m,\omega), \qquad (1)$$

where $S(m,\omega)$ is an unknown source signal at frequency ω. $G_n(\omega)$ is the time-constant acoustic transfer function from the unknown source to the nth microphone. $V_n(m,\omega)$ is the noise at the nth microphone. The N microphone signals can be stacked in a vector format as

$$\mathbf{y}(m,\omega) = \left[\begin{array}{ccc} Y_1(m,\omega) & \cdots & Y_N(m,\omega) \end{array} \right]^T, \qquad (2)$$

where superscript ^{T} denotes the transpose of a vector or matrix.

2.1. LASSO

Let us consider obtaining plane-wave expansion coefficients from a single measurement vector $\mathbf{y}(m,\omega)$. This problem can be formulated with the CS approach [2]. $K(\gg N)$ possible plane-wave incident angles are assumed beforehand. This ap-

proach attempts to obtain a decomposition

$$\mathbf{y}(m,\omega) = \sum_{k=1}^{K} \mathbf{d}_k(\omega) a_k(m,\omega), \tag{3}$$

where $\mathbf{d}_k(\omega)$ $(1 \leq k \leq K)$ is the response of the N sensors of the microphone array to the wave of unit amplitude from kth direction. $a_k(m,\omega)$ is the amplitude and phase of the kth plane wave. This decomposition can be rewritten as

$$\mathbf{y}(m,\omega) = \mathbf{D}(\omega)\mathbf{a}(m,\omega), \tag{4}$$

where

$$\mathbf{D}(\omega) = \left[\begin{array}{ccc} \mathbf{d}_1(\omega) & \cdots & \mathbf{d}_K(\omega) \end{array} \right], \tag{5}$$

$$\mathbf{a}(m,\omega) = \left[\begin{array}{ccc} a_1(m,\omega) & \cdots & a_K(m,\omega) \end{array} \right]^T. \tag{6}$$

$\mathbf{D}(\omega)$ is called a dictionary matrix. $\mathbf{d}_1(\omega) \cdots \mathbf{d}_K(\omega)$ are called atoms of $\mathbf{D}(\omega)$ [9][10]. When some elements have non-zero values and other values are almost zero in the $K \times 1$ vector $\mathbf{a}(m,\omega)$, this sparse $\mathbf{a}(m,\omega)$ is obtained by solving the following l_1 optimization problem called the least absolute shrinkage and selection operator (LASSO) [3][4]

$$\mathbf{a}(m,\omega) = \text{argmin} \\ \| \mathbf{y}(m,\omega) - \mathbf{D}(\omega)\mathbf{a}(m,\omega) \|_2^2 + \lambda \|\mathbf{a}(m,\omega)\|_1, \tag{7}$$

where $\|\mathbf{x}\|_1$ and $\|\mathbf{x}\|_2$ denote the l_1 and l_2 norm of a vector $\mathbf{x}$. λ is a regularization parameter that governs the trade-off between the sparsity of $\mathbf{a}(m,\omega)$ and the fitting error $\mathbf{y}(m,\omega) - \mathbf{D}(\omega)\mathbf{a}(m,\omega)$. λ is the hyper parameter.

2.2. Group LASSO

Next consider estimating plane-wave expansion coefficients from multiple measurement vectors $\mathbf{y}(m,\omega)$ $(1 \leq m \leq M)$. Using multiple measurement vectors is expected to make the estimation more robust to correlated source signals [5][6].

The relationship between the multiple measurement vectors and the corresponding plane-wave expansion coefficients is expressed as

$$\mathbf{Y}(\omega) = \mathbf{D}(\omega)\mathbf{A}(\omega), \tag{8}$$

where

$$\mathbf{Y}(\omega) = \left[\begin{array}{ccc} \mathbf{y}(1,\omega) & \cdots & \mathbf{y}(M,\omega) \end{array} \right], \tag{9}$$

$$\mathbf{A}(\omega) = \left[\begin{array}{ccc} \mathbf{a}(1,\omega) & \cdots & \mathbf{a}(M,\omega) \end{array} \right]. \tag{10}$$

Matrix $\mathbf{A}(\omega)$ can be estimated by applying group LASSO [7]. Group LASSO uses the $l_{1,2}$ norm of $\mathbf{A}(\omega)$ for inducing sparsity. The $l_{1,2}$ norm is defined as

$$\|\mathbf{A}(\omega)\|_{1,2} = \sum_{k=1}^{K} \sqrt{\sum_{m=1}^{M} a_k(m,\omega) a_k^*(m,\omega)}, \tag{11}$$

where $a_k^*(m,\omega)$ is the complex conjugate of $a_k(m,\omega)$. The group LASSO solves the following optimization problem

$$\mathbf{A}(\omega) = \text{argmin } \| \mathbf{Y}(\omega) - \mathbf{D}(\omega)\mathbf{A}(\omega) \|_F^2 + \lambda \|\mathbf{A}(\omega)\|_{1,2}, \tag{12}$$

where $\|\mathbf{X}\|_F$ denotes the Frobenius norm of a matrix $\mathbf{X}$. The $l_{1,2}$ norm acts like LASSO for row vectors of $\mathbf{A}(\omega)$. Depending on λ, some row vectors of $\mathbf{A}(\omega)$ may become almost zero row vector. That is, the $l_{1,2}$ norm selects a few effective atoms. As a result, $\mathbf{Y}(\omega)$ is decomposed to the sum of a few products of $\mathbf{d}_k(\omega)$ and the corresponding row vector in $\mathbf{A}(\omega)$.

2.3. SPICE

SPICE decomposes the PSD matrix $\mathbf{R}(\omega)$ of the microphone signals into the sum of PSD matrices corresponding to sensor noise and to each atom $\mathbf{d}_k(\omega)$ $(1 \leq k \leq K)$ in the dictionary matrix $\mathbf{D}(\omega)$. SPICE reconstructs the PSD matrix as

$$\widehat{\mathbf{R}}(\omega) = \mathbf{D}(\omega) \left[\begin{array}{ccc} \hat{u}_1(\omega) & & 0 \\ & \ddots & \\ 0 & & \hat{u}_K(\omega) \end{array} \right] \mathbf{D}^H(\omega) + \left[\begin{array}{ccc} \hat{v}_1(\omega) & & 0 \\ & \ddots & \\ 0 & & \hat{v}_N(\omega) \end{array} \right]. \tag{13}$$

$\hat{u}_k(\omega)$ $(1 \leq k \leq K)$ is the estimated signal power corresponding to the kth atom $\mathbf{d}_k(\omega)$. $\hat{v}_n(\omega)$ is the estimated noise power at microphone n $(1 \leq n \leq N)$. The noise in the data is taken account of in a natural manner by SPICE. SPICE does not require a hyper parameter such as λ in LASSO and group LASSO.

3. DECOMPOSITION AND ESTIMATION

We propose a method to estimate a reverberant sound field by CS approach without using a hyper parameter. The method first applies SPICE to the PSD matrix of the microphone signals and estimates sensor noise power. The method next builds a convex optimization problem that relates microphone signals, plane-wave expansion coefficients, and sensor noise power in the frequency domain. Its constraint equation is derived using the estimated sensor noise power. The plane-wave expansion coefficients are obtained by solving the convex optimization problem. The sound field can be estimated from the plane-wave expansion coefficients. The entire process of the method is as follows.

Step 1

Compute a PSD matrix $\mathbf{R}(\omega)$ from M measurement vectors $\mathbf{y}(m,\omega)\,(1 \le m \le M)$ as

$$\mathbf{R}(\omega) = \frac{1}{M}\sum_{m=1}^{M} \mathbf{y}(m,\omega)\mathbf{y}^H(m,\omega). \tag{14}$$

Step 2

Decompose $\mathbf{R}(\omega)$ and reconstruct as $\hat{\mathbf{R}}(\omega)$ by applying SPICE as in (13).

Step 3

Estimate vectors $\mathbf{a}_m(\omega)$ $(1 \le m \le M)$ of the plane-wave expansion coefficients by solving the following convex optimization problems

$$\begin{aligned} &\mathbf{a}(m,\omega) = \text{argmin } \|\mathbf{a}(m,\omega)\|_1 \\ &\text{subject to } \|\mathbf{y}(m,\omega) - \mathbf{D}(\omega)\mathbf{a}(m,\omega)\|_2 \le \sqrt{\delta(\omega)} \end{aligned} \tag{15}$$

where

$$\delta(\omega) \quad = \quad \beta(\omega)\left(\hat{v}_1(\omega) + \cdots + \hat{v}_N(\omega)\right). \tag{16}$$

$\beta(\omega)$ is the correction term defined as

$$\beta(\omega) = tr\left(\mathbf{R}(\omega)\right) / tr\left(\hat{\mathbf{R}}(\omega)\right), \tag{17}$$

where $tr(\mathbf{X})$ is the trace of a matrix $\mathbf{X}$. $\beta(\omega)$ makes the sum of microphone signal powers computed from $\hat{\mathbf{R}}(\omega)$ equal to that from $\mathbf{R}(\omega)$. This convex optimization problem (15) is the constrained version of (7) [4].

Step 4

Let $\mathbf{r}$ be the position of a point in the Cartesian coordinate. The sound field due to the kth wave at $\mathbf{r}$ can be estimated from $b_k(m,\omega)$ $(1 \le k \le K,\ 1 \le m \le M)$ by

$$\hat{Y}(\mathbf{r},m,\omega) = \sum_{k=1}^{K} \exp\left(j\frac{\omega}{c}\mathbf{n}_k \bullet \mathbf{r}\right) b_k(m,\omega), \tag{18}$$

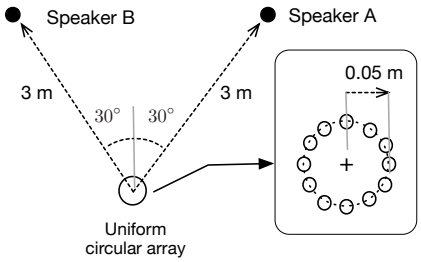

Fig. 1. Arrangement of uniform circular array and sound sources

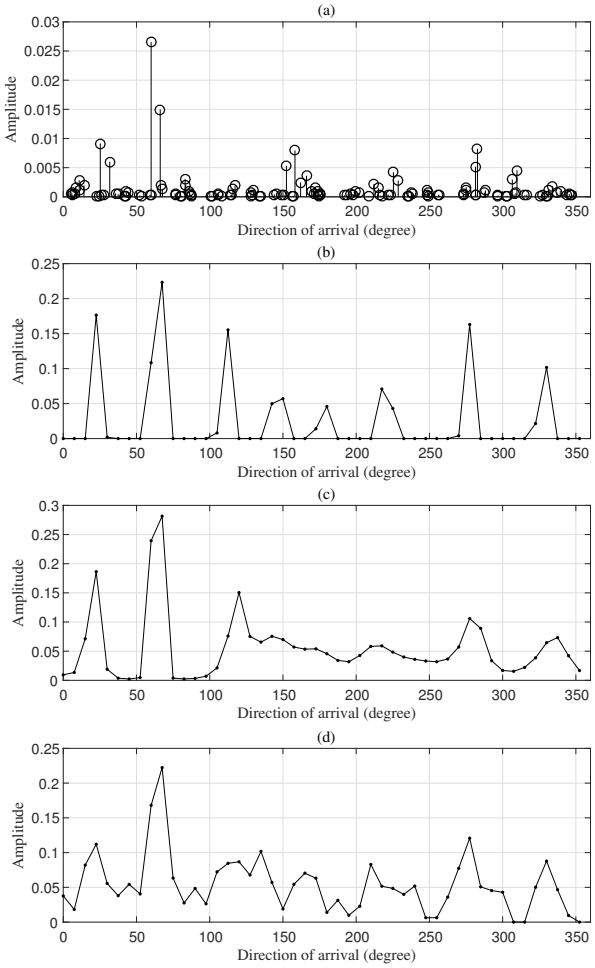

Fig. 2. Direction of arrival of direct and reflected waves and their amplitude: (a) computed from impulse responses generated by image method (true), (b) estimated by group LASSO, (c) estimated by SPICE, and (d) estimated by proposed method.

where $j = \sqrt{-1}$. c is the velocity of sound. $\mathbf{n}_k$ is the unit vector of the direction of the kth plane wave corresponding to $\mathbf{d}_k(\omega)$.

4. EVALUATION

We evaluated the proposed method in a simulated environment. A uniform circular microphone array with a 0.05 m radius and 12 elements was used in a simulated room of $6\times5\times3$ m. The sound sources were positioned 3 m from the circular microphone array as shown in Fig. 1.

All acoustic impulse responses (AIRs) between the sound sources and the microphone array were generated by a modified version [11] [12] of Allen and Berkley's image method [13]. Reflection coefficients of the ceiling and the floor were set to 0. Its reverberation time T_{60} was 320 ms. The sampling frequency was 8 kHz. Microphone signals were gen-

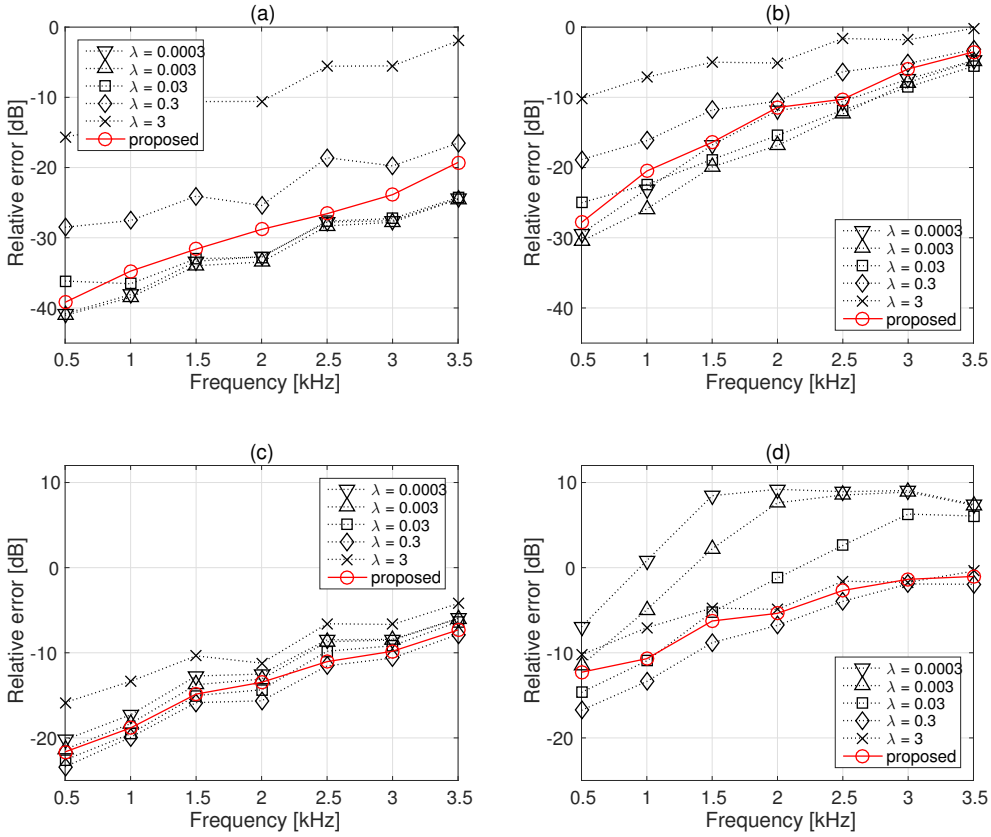

Fig. 3. Relative errors of estimated sound field for group LASSO with various λ and for proposed method (◯): (a) 30 dB SNR, r =0.05 m, (b) 30 dB SNR, r =0.1 m, (c) 10 dB SNR, r =0.05 m, and (d) 10 dB SNR, r = 0.1 m.

erated by convolving the AIRs with pink noise. The frame length of STFT was 512 samples (corresponding to 64 ms). A square-root Hanning window was used. The overlap of STFT was 50%. We assumed 48 possible directions beforehand ($K = 48$ in (5) and (13)). We used 20 frames ($M = 20$). We used SPICE of identical $\hat{v}_n(k)$ [8, III (B)] with iteration number 30. The CVX [14][15] was used for solving the convex optimization problem (15).

First, we evaluated the estimations of the directions of direct and reflected waves when there was only speaker A. The signal-to-noise ratio (SNR) was 30 dB. Fig. 2 shows the results at 2.5 kHz with $\lambda = 0.3$ that was not optimal. Fig. 2 (a) shows the true directions of arrival of direct and reflected waves computed from AIRs generated by the image method. Fig. 2 (b), (c), and (d) show those estimated by group LASSO, SPICE, and the proposed method. The result by group LASSO with non-optimal λ (b) has more spikes than (a). The proposed method (d) has smaller spikes than group LASSO (b).

Second, we evaluated the reconstructed sound field by using twelve points on r =0.05 m and r =0.1 m from the center of the microphone array at the SNRs of 30 dB and 10 dB. The sound field was generated by speaker A and B in Fig. 1. Fig. 3 shows the relative errors for group LASSO with various λ and the proposed method. Fig. 3 (a) and (b) show the relative errors on r =0.05 m and r =0.1 m respectively at 30 dB SNR. Fig. 3 (c) and (d) show those at 10 dB SNR. The relative errors of group LASSO depend on λ. $\lambda = 0.003$ is best at 30 dB SNR and $\lambda = 0.3$ is best at 10 dB SNR. The tuning of the hyper parameter λ is necessary for the group LASSO. Though the relative errors of the proposed method is larger than the lowest relative errors of the group LASSO, the differences are within 6 dB. The proposed method adjusted its estimation for both 30 dB and 10 dB SNR without tuning.

5. CONCLUSION

We have developed a method to estimate a reverberant sound field by compressed sensing approach without using a hyper parameter for controlling sparsity. The proposed method was shown to adjust its sound field estimation for both at 30 dB SNR and 10 dB SNR without tuning.

6. REFERENCES

[1] E. G. Williams, *Fourier Acoustics*, Academic, New York, 2000.

[2] P. K. T. Wu, N. Epain, and C. Jin, "A dereverberation algorithm for spherical microphone arrays using compressed sensing techniques," in *Proc. ICASSP2012*, 2012, pp. 4053–4056.

[3] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Rev.*, vol. 43, no. 1, pp. 129–159, 2001.

[4] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer, New York, 2001.

[5] D. Malioutov, M. Cetin, and A. S. Willsky, "A sparse signal reconstruction perspective for source localization with sensor arrays," *IEEE Trans. Signal Process.*, vol. 53, no. 8, pp. 3010—3022, 2005.

[6] S. F. Cotter, B. Rao, K. Engan, and K. Kreutz-Delgado, "Sparse solutions to linear inverse problems with multiple measurement vectors," *IEEE Trans. Signal Process.*, vol. 53, no. 7, pp. 2477–2488, 7 2005.

[7] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B*, vol. 68, no. 1, pp. 49—67, 2006.

[8] P. Stoica, P. Babu, and J. Li, "Spice: A sparse covariance-based estimation method for array processing," *IEEE Trans. Signal Process.*, vol. 59, no. 2, pp. 629–638, Feb. 2011.

[9] E. J. Cande's and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 21–30, 2008.

[10] R. Rubinstein, A. M. Bruckstein, and M. Elad, "Dictionaries for sparse representation modeling," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 1045–1057, 2010.

[11] P. M. Peterson, "Simulating the response of multiple microphones to a single acoustic source in a reverberant room," *J. Acoust. Soc. Amer.*, pp. 1527—1529, Nov. 1986.

[12] E. Habets, "Room impulse response (rir) generator," http://home.tiscali.nl/ehabets/rirgenerator.html, July 2006.

[13] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small room acoustics," *J. Acoust. Soc. Amer.*, pp. 943—950, 1979.

[14] Inc. CVX Research, "CVX: Matlab software for disciplined convex programming," http://cvx.com/cvx, Aug. 2012.

[15] M. Grant and S. Boyd, "Graph implementations for nonsmooth convex program," in *Recent Advances in Learning and Control*, V. Blondel, S. Boyd, and H. Kimura, Eds., Lecture Notes in Control and Information Sciences, pp. 95–100. Springer-Verlag, 2008.