BAYESIAN REGULARIZATION AND NONNEGATIVE DECONVOLUTION (BRAND) FOR ACOUSTIC ECHO CANCELLATION

*Yuanqing Lin and Daniel D. Lee**

GRASP Laboratory, Department of Electrical and Systems Engineering
University of Pennsylvania, Philadelphia, PA 19104
linyuanq, ddlee@seas.upenn.edu

ABSTRACT

The Bayesian Regularization And Nonnegative Deconvolution (BRAND) algorithm is used to estimate the acoustic room impulse response by incorporating prior information such as sparsity and nonnegativity about the filter coefficients. For experimental measurements with microphones and speakers with non-ideal characteristics, the overall transfer function can be decomposed into a common short FIR filter, and a long nonnegative filter representing the room response. We develop an online estimation procedure for the BRAND algorithm, along with a computationally efficient implementation. Simulations and experimental results show the robustness of the resulting algorithm for echo cancellation in the presence of large ambient noise.

1. INTRODUCTION

Acoustic echo cancellation [1] is crucial for many applications such as hands-free telephony and teleconferencing. However, efficient and robust estimation of the room impulse response becomes challenging in environments with long filter responses and large ambient noise. Most conventional algorithms based upon least mean square (LMS) procedures [2] may not properly converge in the presence of such noise. When long delays with a large number of filter coefficients need to be estimated, computational efficiency becomes important for real-time applications. Another difficulty with echo cancellation is that the impulse response may change over time, as the source or microphone is moved.

Some of these difficulties are alleviated by incorporating prior knowledge about the acoustic room impulse response in the echo cancellation algorithm. Our recently proposed Bayesian Regularization and Nonnegative Deconvolution (BRAND) algorithm uses the sparseness and nonnegativity of theoretical room impulse responses to robustly and efficiently estimate room impulse coefficients even in the presence of a large amount of noise ($>$ -20 dB) [3]. Our previous work has demonstrated how the algorithm can be used in batch mode; in this work, we develop an online version of the update rules for the BRAND algorithm that can be efficiently implemented using FFT's.

Although the theoretical room impulse response may be nonnegative and sparse [4], experimental measurements of the transfer function between a source signal $s_0(t)$ and a measured signal $x(t)$ may not exhibit these properties. This is due to the non-ideal characteristics of the speaker, microphone, and intervening sound hardware. To compensate for these effects, we model the overall impulse response as the convolution of a theoretical room impulse response with the transfer function of the sound hardware:

$$x(t) = s_0(t) * h(t) * \alpha(t) \tag{1}$$

where $h(t)$ is a short FIR filter describing the impulse response of the sound hardware, $\alpha(t)$ is a long nonnegative filter representing the room impulse response, and $*$ denotes the convolution operation. Since $h(t)$ is fixed for a given set of sound hardware, it can be measured or estimated beforehand. Then, the system identification task in echo cancellation is reduced to estimating $\alpha(t)$ using the appropriate sparse, nonnegative priors. The optimization problem for estimating $\alpha(t)$ becomes:

$$\min_{\alpha(t)\geq 0} \int dt \, \frac{1}{2} \left\| x(t) - \int dt' \alpha(t') s(t-t') \right\|^2 + \lambda(t)\alpha(t). \tag{2}$$

where $s(t) = s_0(t) * h(t)$, and $\lambda(t) \geq 0$ is the L_1-norm sparsity regularization function.

The sound hardware impulse response $h(t)$ could be measured in an anechoic chamber. However, here we introduce an algorithm that estimates $h(t)$ using a number of measurements taken in a normal environment. This algorithm is similar to the one that we proposed for source estimation [5]. The optimization problem in discrete form is

$$\min_{\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_K\geq 0,\mathbf{h}} \sum_{k}^{K} \|\mathbf{x}_k - \mathbf{s}_{0k} * \mathbf{h} * \boldsymbol{\alpha}_k\|^2 + \sum_{k=1}^{K}\sum_{i=1}^{M} \hat{\lambda}_{ki}\alpha_{ki}. \tag{3}$$

where $\mathbf{x}_k = [x_k(1), x_k(2), ..., x_k(N)]^T$ is an N$\times$1 vector representing the k^{th} signal detected by the microphone, $\mathbf{s}_{0k}$ the k^{th} source signal, $\boldsymbol{\alpha}_k$ is a M$\times$1 vector describing the room impulse response associated with the k^{th} position of the speaker-microphone pair, $\mathbf{h}$ is an L$\times$1 vector, and $\hat{\lambda}_{ki}$ is the L_1-norm sparsity regularization parameter that penalizes non-zero solutions of α_{ki} [6]. This optimization problem can be solved by alternatively iterating the two steps: α-step and h-step. In the α-step, the nonnegative filter coefficients $\{\boldsymbol{\alpha}_k\}_{k=K1}^{K}$ are estimated with respect to the current $\mathbf{h}$ estimate using the BRAND algorithm to infer the optimal regularization parameters and sparse solutions. Then, in the h-step, $\mathbf{h}$ is optimized with respect to the new $\{\boldsymbol{\alpha}_k\}_{k=1}^{K}$ estimates.

After the FIR filter $\mathbf{h}$ associated with the sound hardware is estimated, the optimization problem for estimating the room impulse response with nonnegativity and sparsity constraints becomes

$$\min_{\boldsymbol{\alpha}\geq 0} \frac{1}{2\sigma^2}\|\mathbf{x} - \mathbf{S}\boldsymbol{\alpha}\|^2 + \sum_i \lambda_i \alpha_i. \tag{4}$$

*This work was funded by the U.S. Army Research Office

where σ^2 and λ_i ($\sigma^2\lambda_i = \hat{\lambda_i}$, as shown in Eq. 3) are sparsity regularization parameters that is inferred by the BRAND algorithm, $\mathbf{x}$ is the discrete sequence

$$\mathbf{x} = [x(i-N+1), x(i-N+2), ..., x(i)]^T;$$

where i is time index, and $\mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_M]$ is an N×M matrix whose columns are the delayed patterns of the source $s(t)$ ($= s_0(t) * h(t)$), so that $\mathbf{S}$ has the form:

$$\begin{bmatrix} s(i-N+1) & s(i-N) & \dots & s(i-N-M+2) \\ s(i-N+2) & s(i-N+1) & \dots & s(i-N-M+3) \\ \vdots & \vdots & \vdots & \vdots \\ s(i) & s(i-1) & \dots & s(i-M+1) \end{bmatrix}. \quad (5)$$

For online echo cancellation, the algorithm is applied to moving windows of measured signal and the estimate of the room impulse response is updated iteratively. For these updates, we exploit the Toeplitz structure in the $\mathbf{S}$ matrix for computational efficiency using FFT's.

The remainder of this paper is organized as follows. Section 2 describes the algorithm consisting of the α-step and h-step for estimating the short FIR filter $h(t)$ associated with the sound hardware. The α-step employs the BRAND algorithm, and a fast implementation for the multiplicative update is introduced for solving the associated nonnegative quadratic programming problem. Section 3.1 presents the results of estimating the filter $h(t)$ using 32 experimental measurements using generic sound hardware, as well as the results of cross-validating the resulting $h(t)$ estimate. In Section 3.2, we employ simulations to quantitatively compare the on-line performance of BRAND and NLMS with very noisy data, and Section 3.3 shows the echo cancellation performance of the online BRAND algorithm in an actual acoustic environment. Finally, we conclude with a discussion of these results in Section 4.

2. ALGORITHM FOR ESTIMATING *h(t)*

In this section, we describe the algorithm for estimating the FIR filter $h(t)$ associated with the sound hardware from a number of acoustic measurements using the optimization in Eq. 3. It is solved by alternately iterating two steps: an α-step and a h-step. In the α-step, given the current estimate of $\mathbf{h}$, the BRAND algorithm is independently applied to each data set and infers the sparsity regularization parameters $\hat{\boldsymbol{\lambda}}_k$ as well as the nonnegative estimates $\boldsymbol{\alpha}_k$, for $k = 1, 2, .., K$. Here we only review the update rules for the algorithm, and more details about the underlying Bayesian framework can be found in [3]. In the h-step, the FIR filter $\mathbf{h}$ is iteratively optimized by solving an unconstrained LMS problem.

We also introduce a FFT based implementation of the BRAND algorithm. The resulting algorithm is very efficient in terms of both computational load and memory usage, and will be used for adaptively estimating the room impulse response for echo cancellation.

2.1. α-step: Bayesian Regularization And Nonnegative Deconvolution (BRAND)

The BRAND algorithm finds nonnegative solutions of Eq. 4 with the appropriate sparseness by inferring the sparsity regularization parameters (σ^2 and $\boldsymbol{\lambda}$) in a Bayesian framework. Its EM-like updates alternatively iterates solving a nonnegative quadratic programming (NNQP) problem for the most likely $\boldsymbol{\alpha}$, denoted as $\boldsymbol{\alpha}^{ML}$, with respect to the current regularization parameters, and then re-estimating the regularization parameters from the latest $\boldsymbol{\alpha}^{ML}$ estimate.

The NNQP problem in Eq. 4 can be put in standard form:

$$\min_{\boldsymbol{\alpha} \geq 0} \frac{1}{2}\boldsymbol{\alpha}^T\mathbf{A}\boldsymbol{\alpha} + \mathbf{b}^T\boldsymbol{\alpha} \quad (6)$$

with $\mathbf{A} = \frac{1}{\sigma^2}\mathbf{S}^T\mathbf{S}$ and $\mathbf{b} = -\frac{1}{\sigma^2}\mathbf{S}^T\mathbf{x} + \boldsymbol{\lambda}$. There are various methods for solving the NNQP problem, including the simplex method and interior point methods. We employ multiplicative updates for solving Eq. 6 which are particularly easy to implement and have guaranteed convergence properties [7]:

$$\alpha_i \longleftarrow \alpha_i \frac{-b_i + \sqrt{b_i^2 + 4(\mathbf{A}^+\boldsymbol{\alpha})_i(\mathbf{A}^-\boldsymbol{\alpha})_i}}{2(\mathbf{A}^+\boldsymbol{\alpha})_i}, \quad (7)$$

where

$$A_{ij}^+ = \begin{cases} A_{ij} & \text{if} \quad A_{ij} > 0 \\ 0 & \text{if} \quad A_{ij} \leq 0 \end{cases} \quad A_{ij}^- = \begin{cases} 0 & \text{if} \quad A_{ij} \geq 0 \\ -A_{ij} & \text{if} \quad A_{ij} < 0 \end{cases}. \quad (8)$$

From the estimated $\boldsymbol{\alpha}^{ML}$, the regularization parameters are re-estimated as:

$$\lambda_i \longleftarrow \frac{1}{\bar{\alpha}_i} \quad (9)$$

$$\sigma^2 \longleftarrow \frac{1}{N}[(\mathbf{x} - \mathbf{S}\bar{\boldsymbol{\alpha}})^T(\mathbf{x} - \mathbf{S}\bar{\boldsymbol{\alpha}}) + \text{Tr}(\mathbf{S}^T\mathbf{S}\mathbf{C})] \quad (10)$$

where $\bar{\alpha}$ and $\mathbf{C}$ represent the mean and covariance of a variational approximation for the distribution of $\boldsymbol{\alpha}$:

$$\bar{\alpha}_i = \begin{cases} \alpha_i^{ML} & \text{if} \quad i \in J \\ \mu_i & \text{if} \quad i \in I \end{cases} \quad (11)$$

$$C_{ij} = \begin{cases} (\mathbf{A_{JJ}}^{-1})_{ij} & \text{if } i, j \in J \\ \mu_i^2\delta_{ij} & \text{otherwise,} \end{cases} \quad (12)$$

I and J are the indices of zero and non-zero elements in $\boldsymbol{\alpha}^{ML}$, respectively, while $\boldsymbol{\mu}$ are variational parameters that model the distribution of the zero elements of $\boldsymbol{\alpha}^{ML}$ obtained from:

$$\mu_i \longleftarrow \mu_i \frac{-\hat{b}_i + \sqrt{\hat{b}_i^2 + 4(\hat{\mathbf{A}}^+\boldsymbol{\mu})_i[(\hat{\mathbf{A}}^-\boldsymbol{\mu})_i + \frac{1}{\mu_i}]}}{2(\hat{\mathbf{A}}^+\boldsymbol{\mu})_i}. \quad (13)$$

where $\hat{\mathbf{b}}_I = (\mathbf{A}\boldsymbol{\alpha}^{ML} + \mathbf{b})_I$, $\hat{\mathbf{A}} = \mathbf{A}_{II} + \text{diag}(\mathbf{A}_{II})$ with $\hat{\mathbf{A}} = \hat{\mathbf{A}}^+ - \hat{\mathbf{A}}^-$ being the decomposition of $\hat{\mathbf{A}}$ into its positive and negative components.

2.1.1. Fast implementation of the multiplicative updates

In typical acoustic echo cancellation applications, the dimensionality of $\boldsymbol{\alpha}$ is often larger than 1000. As a result, both computational load and memory usage can be quite costly for the matrix-vector multiplications, $\mathbf{A}^+\boldsymbol{\alpha}$, $\mathbf{A}^-\boldsymbol{\alpha}$, $\hat{\mathbf{A}}^-\boldsymbol{\mu}$, and $\hat{\mathbf{A}}^+\boldsymbol{\mu}$, in Eqs. 7 and 13. Instead of directly computing these terms, we exploit the Toeplitz structure of matrix $\mathbf{S}$ in Eq. 5. Note that the decomposition of $\mathbf{A}$ into $\mathbf{A}^+$ and $\mathbf{A}^-$ is arbitrary as long as the elements of these two matrices are nonnegative. We then decompose $\mathbf{A} = \mathbf{A}^+ - \mathbf{A}^-$ in terms of $\mathbf{S}^+$ and $\mathbf{S}^-$, so that $\mathbf{S} = \mathbf{S}^+ - \mathbf{S}^-$ which is similar to Eq. 8, namely

$$\begin{aligned} \mathbf{A}^+ &= \frac{1}{\sigma^2}(\mathbf{S}^{+T}\mathbf{S}^+ + \mathbf{S}^{-T}\mathbf{S}^-), \\ \mathbf{A}^- &= \frac{1}{\sigma^2}(\mathbf{S}^{+T}\mathbf{S}^- + \mathbf{S}^{-T}\mathbf{S}^+). \end{aligned} \quad (14)$$

With this decomposition, $\mathbf{A}^+\boldsymbol{\alpha}$ and $\mathbf{A}^-\boldsymbol{\alpha}$ in Eq. 7 can be implemented by FFT's since both $\mathbf{S}^+$ and $\mathbf{S}^-$ are Toeplitz matrices. The computations $\hat{\mathbf{A}}^-\boldsymbol{\mu}$ and $\hat{\mathbf{A}}^+\boldsymbol{\mu}$ are implemented in the same way except that $\boldsymbol{\mu}$ needs to be zero-padded appropriately. Although the decomposition in Eq. 14 may slow convergence compared to Eq. 8, it dramatically improves computational efficiency. The resulting algorithm is suitable for real-time implementation in acoustic echo cancelling systems.

2.2. h-step

Given the latest estimates of $\boldsymbol{\alpha}_k$ (k=1,2,...,K), let $\mathbf{s}_{rk} = \mathbf{s}_{0k} * \boldsymbol{\alpha}_k$, and $\mathbf{S}_{rk}$ denote the matrix whose columns are delayed versions of $\mathbf{s}_{rk}$. Then the optimization problem in Eq.4 with respect to h has an analytic solution, which can be written in closed form:

$$h = (\sum_k^K \mathbf{S}_{rk}^T \mathbf{S}_{rk})^{-1} (\sum_k^K \mathbf{S}_{rk}^T \mathbf{x}_k) \qquad (15)$$

3. RESULTS

In this section, we show and evaluate the resulting estimates of the FIR filter $h(t)$ associated with a speaker-microphone pair using signals acquired from 32 different room locations. Next, we use simulations to quantitatively illustrate the performance of the on-line BRAND algorithm in comparison with that of the NLMS algorithm. Finally, we demonstrate the echo cancellation performance of the on-line BRAND algorithm in real environments using the estimated $h(t)$.

3.1. $h(t)$ estimation

In order to estimate the FIR filter ($h(t)$) describing the transfer function of a speaker-microphone pair, we used this pair to acquire data sets at 32 different locations, each of them was associated with a different room impulse response ($\alpha(t)$). For each data set, a random white pseudonoise analog signal was fed to the soundcard to drive the speaker. Then, both this analog signal ($s_0(t)$) and the detected signal on the microphone ($x(t)$) were recorded by a stereo recorder, similar to typical echo cancellation hardware. Since our goal is to estimate $h(t)$, the length of the room impulse responses $\alpha(t)$ was typically limited to less than 64ms (1024 taps for a sampling rate of 16kHz).

The signals $\mathbf{s}_{0k}$ and $\mathbf{x}_k$ ($k = 1, 2, ..., 32$) are the first 4096 samples of the recorded speaker and microphone signals, respectively. To avoid anti-aliasing effects, a 48-order FIR bandpass filter was applied to the microphone signals, denoted by $h_B(t)$. Consequently, the estimated $\hat{h}(t)$ (128 taps) is the combination of the impulse response of the speaker-microphone pair, $h(t)$, and the designed bandpass filter, $h_B(t)$. In the α-step, from the current $\hat{h}(t)$ estimate, BRAND as described in Section 2.1 is applied independently on each data set to find the nonnegative estimates of the room impulse responses, namely, $\boldsymbol{\alpha}_k$, $k = 1, 2, ..., 32$. Then from these $\boldsymbol{\alpha}$ estimates, $\hat{h}(t)$ is further optimized in the h-step using the LMS update described in Eq. 15. By alternating the α-step and the h-step, the algorithm gives the final $\hat{h}(t)$ estimate shown in Fig. 1 (b).

The resulting $\hat{h}(t)$ estimate could be compared to a direct anechoic measurement; instead, we validate the estimate in the following manner. The average residual error of the fits over the 32 data sets was -24.8dB compared to an average of -29.7dB when an unconstrained LMS was directly applied to $s_{0k}(t)$ and $x_k(t) * h_B(t)$ with $k = 1, 2, ..., 32$, indicating that the assumptions of nonnegativity and sparsity properties are appropriate to explain the data. We also performed a cross-validation experiment to demonstrate consistency. An additional measurement was taken in a reverberant small room which had echoes as long as 200ms (3072 taps with sampling rate of 16kHz). Fig. 1(a) shows the resulting filter estimate by an unconstrained LMS fit between the filtered microphone signal $x(t) * h_B(t)$ and the source $s_0(t)$. We also show that this filter can alternatively be decomposed into a short FIR filter $\hat{h}(t)$, shown in Fig. 1(b), and a long nonnegative filter, shown in Fig. 1(c), with residual error of -25.2dB compared to the unconstrained fit with residual -29.2dB. As shown in Fig. 1(d), the resulting predictions on a speech signal using this decomposition is quite similar to the unconstrained impulse response filter.

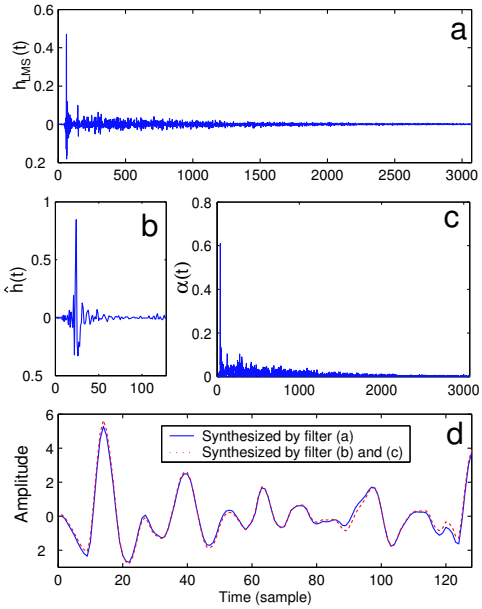

Figure 1: Estimated $\hat{h}(t)$ and its cross-validation. (a) Unconstrain LMS solution ($h_{LMS}(t)$) with fitting error -29.2dB; (b) Estimated $\hat{h}(t)$, (c) BRAND result ($\alpha(t)$) with fitting error -25.2dB, it has about 340 elements less than 10^{-5}, (d) The convolution results (only shows a short segment) of a speech respectively with $h_{LMS}(t)$ and $\hat{h}(t) * \alpha(t)$.

3.2. Performance comparison between BRAND and NLMS using simulations

We have verified that the overall impulse response can be decomposed into two filters: one short FIR filter associated with the speaker-microphone pair, and one long nonnegative filter describing the particular room impulse response. In this section, simulations are employed to quantitatively illustrate the advantage of this approach.

For the simulations, we employed a 18-second speech signal as the source $s(t)$. The simulated microphone signal $x(t)$ was the convolution of the source and the simulated room impulse response, which is the first 2048 taps of the filter shown in Fig. 1 (c).

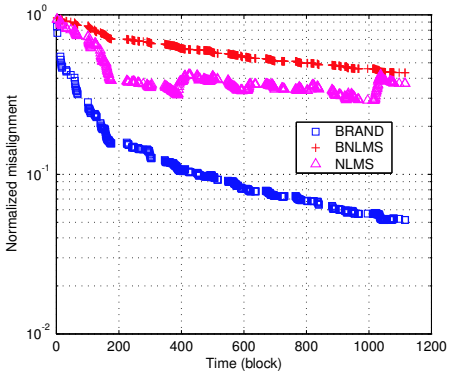

Figure 2: Normalized misalignment $\frac{\|\boldsymbol{\alpha}-\boldsymbol{\alpha}_0\|^2}{\|\boldsymbol{\alpha}_0\|^2}$ ($\boldsymbol{\alpha}$, the estimated filter, and $\boldsymbol{\alpha}_0$, the true filter) by three different algorithms: BRAND, normalized least-mean-square (NLMS), and block NLMS (BNLMS). The microphone signal was corrupted by -10dB Gaussian white noise.

The microphone signal was corrupted by -10dB Gaussian white noise.

Fig. 2 shows the normalized misalignment, $\frac{\|\boldsymbol{\alpha}-\boldsymbol{\alpha}_0\|^2}{\|\boldsymbol{\alpha}_0\|^2}$ with $\boldsymbol{\alpha}$ being the estimated filter and $\boldsymbol{\alpha}_0$ being the actual one, by three different algorithms: BRAND, normalized least-mean-square (NLMS), and block NLMS (BNLMS). The signals were split into windows of 1024 samples with an overlap of 768 samples between neighboring windows. A silence detection algorithm was employed to decide if a particular window would be used for adaption or not. In simulation, both BRAND algorithm and BNLMS algorithm were implemented efficiently by FFT's, and 10 iterations were used. On our Pentium 4 platform, the computation of BRAND, BNLMS and NLMS for a window took 0.43, 0.13, and 0.02 seconds, respectively. Our results also showed that nonnegative deconvolution without Bayesian regularization would lead to a similar result as BRAND when the ambient noise is less than -20dB, and the simplified algorithm will cut the computation in half.

3.3. Echo cancellation in real environments

To test the performance of the BRAND algorithm with the estimated FIR filter $\hat{h}(t)$ in a real acoustic environment, we acquired two sets of measurements using different 20-second speech segments as the source but taken in the same room location. As a result, these two sets of measurement shared the same room impulse response. We used one set of signals, denoted the training set, to estimate the room impulse response in an online fashion, and the resulting estimate was used to perform echo cancellation on the second test set. To illustrate the robustness of the BRAND algorithm to ambient noise, the signals in the training set was corrupted by -10dB white Gaussian noise. The resulting echo cancellation performance is shown in Fig. 3. It indicates that the echo cancellation algorithm obtained attenuations between 10 to 20 dB during the large energy portions of the signal.

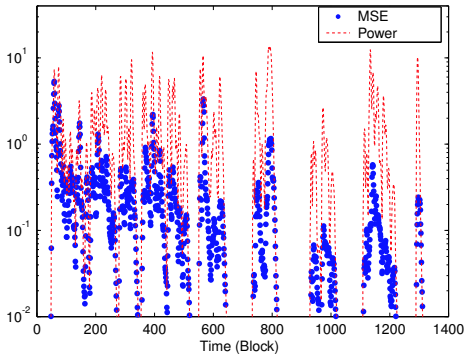

Figure 3: The mean square error (MSE) of echo cancellation (dot) in comparison with the power of the signal (dash line). The microphone signal in the training set was corrupted by -10dB noise.

4. DISCUSSION

We have demonstrated that the overall room impulse response for echo cancellation can be decomposed into a short FIR filter associated with the sound hardware, and a sparse, nonnegative long filter that describes the acoustic room impulse response. The constraints associated with the nonnegative filter make it highly advantageous for echo cancellation applications, since its estimation is very robust to ambient noise. Furthermore, an efficient implementation using FFT's for these new algorithms makes possible their use in real-time acoustic signal processing systems.

5. REFERENCES

[1] J. Benesty, T. Gansler, D. R. Morgan, M. M. Sondhi, and S. L. Gay, *Advances in Network and Acoustic Echo Cancellation*. Springer-Verlag, 2001.

[2] G. Glentis, K. Berberidis, and S. Theodoridis, "Efficient least squares adaptive algorithms for FIR transversal filtering," *IEEE Signal Processing Magazine*, vol. 16, pp. 13–41, 1999.

[3] Y. Lin and D. D. Lee, "Bayesian Regularization And Nonnegative Deconvolution (BRAND) for room impulse response estimation," *IEEE Trans. Signal Processing*, Accepted for publication.

[4] 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.

[5] Y. Lin and D. D. Lee, "Relevant deconvolution for acoustic source estimation," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2005.

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

[7] F. Sha, L. K. Saul, and D. Lee, "Multiplicative updates for nonnegative quadratic programming in support vector machines," in *Advances in Neural Information Processing Systems*, S. T. Suzanna Becker and K. Obermayer, Eds., vol. 15. The MIT Press, 2002.