BLIND DECONVOLUTION OF REVERBERATED SPEECH SIGNALS VIA REGULARIZATION

Juan Liu

Univ. of Illinois, Beckman Institute
405 N. Mathews Ave., Urbana, IL 61801

Henrique Malvar

Microsoft Research
One Microsoft Way, Redmond, WA 98052

ABSTRACT

This paper explores blind deconvolution of reverberated speech signals in microphone array applications. Two regularization approaches are proposed based on available *a priori* knowledge. The regularized least-squares (LS) approach uses the speech signal characteristics and the lowpass nature of the reverberation channel; and the regularized cross correlation (CR) approach requires more precise knowledge of reverberation which can be obtained through training. The two methods are robust to the presence of noise.

1. INTRODUCTION

In a typical office room environment, speech signals are distorted due to the reflection of walls, whiteboards, furniture, and other objects. The signal recorded at a microphone or a microphone array often sounds reverberated because of the multipath effect. The reverberation brings difficulty to speech processing such as recognition and compression. Recovering the clean signal from the reverberated observation is thus an important problem.

This paper discusses blind deconvolution of reverberated speech signals in microphone array applications. Fig. 1 describes a microphone array system with P microphones. The observation $\underline{y}_i$ $(1 \leq i \leq P)$ denotes the signal recorded at the ith microphone. Each $\underline{y}_i$ is a reverberated version of the original clean speech $\underline{x}$, and the channel $h_i(n)$ characterizes the acoustic distortion from the sound source to the ith microphone. In practical microphone array systems, channels are often time-varying due to the movement of the sound source and changes in the environment. The goal is to recover the clean signal $\underline{x}$ and the channels $h_i(n)$ from the reverberated signal $y_i(n)$ $(1 \leq i \leq P)$.

One important issue in blind deconvolution is identifiability: can the signal and the channels be identified uniquely (up to scaling factors) from the outputs? Xu *et al* in [1] and Hua and Wax in [2] give identifiability conditions for multiple channel blind deconvolution applications: first, the channels should be coprime, i.e., they cannot share common zeros; second, the signal $\underline{x}$ must have sufficient spectral richness. Based on those analyses, eigenstructure-based or likelihood-based methods have been proposed. See [1, 3, 4, 5, 6] for examples.

In practical microphone array systems, signals are often corrupted by noise, which complicates the blind deconvolution problem. Hua [4] gives a comparison of representative blind deconvolution methods and also the Cramer-Rao lower bound. The performance of such identification methods decreases quickly as the signal-to-noise ratio (SNR) decreases (e.g., $\leq$ 30dB). This calls for suitable *a priori* knowledge to stabilize the solution, and to restrict the size of the solution set. This paper presents two such regularization approaches.

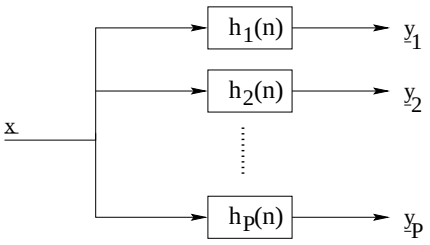

Fig. 1. A multichannel system

2. A REGULARIZED LS APPROACH

The basic idea is to penalize unlikely estimates by a regularization penalty. The choice of the penalty is often critical to the performance. In this section, we present a blind deconvolution approach using the characteristics of speech signal and the lowpass nature of reverberation channels. We define the composite observation vector $\underline{y} \stackrel{\triangle}{=} (\underline{y}_1^T, \underline{y}_2^T, \ldots, \underline{y}_P)^T$, and the impulse response vector $\underline{h} \stackrel{\triangle}{=} (\underline{h}_1^T, \underline{h}_2^T, \ldots \underline{h}_P^T)^T$. Let H_i be the Toeplitz convolution matrix for the ith channel, and let

$H = (H_1^T, H_2^T, \ldots, H_P^T)^T$ be the composite convolution operator. We design the cost function to be minimized as

$$\mathcal{E} = ||\underline{y} - H\underline{x}||^2 + \mu_x \Phi_x(\underline{x}) + \mu_h \Phi_h(\underline{h}). \qquad (1)$$

where each term has the following interpretation:

- $||\underline{y} - H\underline{x}||^2$ is a least-squares metric, the L^2 distance between the data $\underline{y}$ and the channel outputs given by the estimate $(\underline{x}, \underline{h})$. It measures the fidelity to the data.
- $\Phi_x(\underline{x})$ is the regularization penalty on $\underline{x}$, penalizing unlikely signals. In a microphone array system, $\underline{x}$ is mostly speech, which can be well predicted by linear predictive coding (LPC) analysis. Hence we choose the LPC residual energy as the regularization penalty. More specifically, $\Phi_x(\underline{x}) = ||C_x\underline{x}||^2$, where C_x denotes the Toeplitz convolution operator obtained from the LPC coefficients.
- $\Phi_h(\underline{h})$ penalizes unlikely room impulse responses. The reverberation components of room impulse responses are usually lowpass, because high-frequency sound waves are more likely to be absorbed by walls and other surfaces. We choose $\Phi_h(\underline{h})$ to be the total energy of the output of convolving each $\underline{h}_i$ with a simple 4-tap highpass filter $c = \{1, -1, 0.6, -0.3\}$. This simple filter has worked well in our experiments.
- The regularization parameters μ_x and μ_h control the tradeoff between the least-squares term and the regularization penalties.

2.1. Optimization Algorithm

Finding the $(\hat{\underline{h}}, \hat{\underline{x}})$ that minimizes $\mathcal{E}$ is a high-dimensional nonlinear optimization problem. We solve it by a coordinate descent approach [7], which takes turns in optimizing with respect to $\underline{x}$, $\underline{h}$, and C_x. After proper initialization, the algorithm does the following:

- Fix $\underline{h}$ and C_x, and find $\underline{x} = \arg\min_{\underline{x}} \{||\underline{y} - H\underline{x}||^2 + \mu_x ||C_x\underline{x}||^2\}$. Since the H_i and C_x are all convolution operators, the optimization problem can be solved using frequency domain filtering and inverse filtering operations.
- Fix $\underline{x}$ and find C_x. This is equivalent to finding the LPC coefficients of $\underline{x}$.
- Fix $\underline{x}$, find $\underline{h} = \arg\min_{\underline{h}} \left\{ ||\underline{y} - H\underline{x}||^2 + \mu_h \sum_{i=1}^{P} \Phi(\underline{h}_i) \right\}$. Each channel can be estimated separately.

The steps above are repeated in sequence, until convergence is attained. Each step is a quadratic optimization problem, for which a global minimum is guaranteed to be found. Therefore, the algorithm is guaranteed to converge to a local minimum [7].

In practice, to further reduce the computational cost, the observation is divided into blocks of suitable length (e.g. 32 ms). The algorithm iteratively finds the estimate, and carries the estimated channel to the succeeding blocks. Moreover, we use a forgetting factor to stabilize $\underline{h}$, i.e., $\underline{h}^{k+1}_{initial} = (1-\alpha)\underline{h}^k_{initial} + \alpha\underline{h}^k_{estimate}$, where k is the block index, and $\underline{h}^k_{initial}$ and $\underline{h}^k_{estimate}$ denote the initialized value and the optimized estimate in block k respectively. The forgetting factor $\alpha \in (0, 1)$ controls the learning rate of $\underline{h}$, prohibiting it to vary too quickly. In essence, noise in successive blocks are smoothed by using small forgetting factors. In our experiments, we use α between 0.1 and 0.2. As more blocks are processed, the estimate of $\underline{h}$ gets more accurate, i.e. the blind deconvolution algorithm learns the reverberation channels gradually.

2.2. Experimental Results

To evaluate the algorithm proposed above, we generated the observations $\underline{y}$ by passing a signal $\underline{x}$ through five 80-tap FIR channels $h_1(n), h_2(n), \ldots, h_P(n)$ (with $P = 5$ and all filters having length $L = 80$). Each channel is contaminated with white noise with an SNR of 24 dB. Figure 2 shows an of the performance of our regularized least-squares approach. The estimated channel after processing 40,000 samples (the dashed curve) is close to its true value (the solid curve), both in the time domain (see Fig. 2a) and in the frequency domain (see Fig. 2b). The SNR in estimating $|H(\omega)|$ is 9.7 dB.

We compared the restored speech signal $\hat{x}$ using our technique to the result using a simple delay-and-sum beamforming technique (in which the multiple channel observations $\underline{y}_1, \underline{y}_2, ..., \underline{y}_P$ are time aligned and then averaged [8]). The SNR gain of our $\hat{x}$ over delay-and-sum beamforming estimate is 5.6 dB. Similar results are achieved when the channels are contaminated with nonwhite or ambient noise, which suggests the robustness of the regularized least-squares method.

3. A REGULARIZED CR APPROACH

In a system with $P = 2$ channels, it is easy to see that in the absence of noise, we have $\underline{y}_1 * \underline{h}_2 - \underline{y}_2 * \underline{h}_1 = x * (h_1 * h_2 - h_2 * h_1) = 0$. This observation forms the basis of cross-correlation blind identification methods.

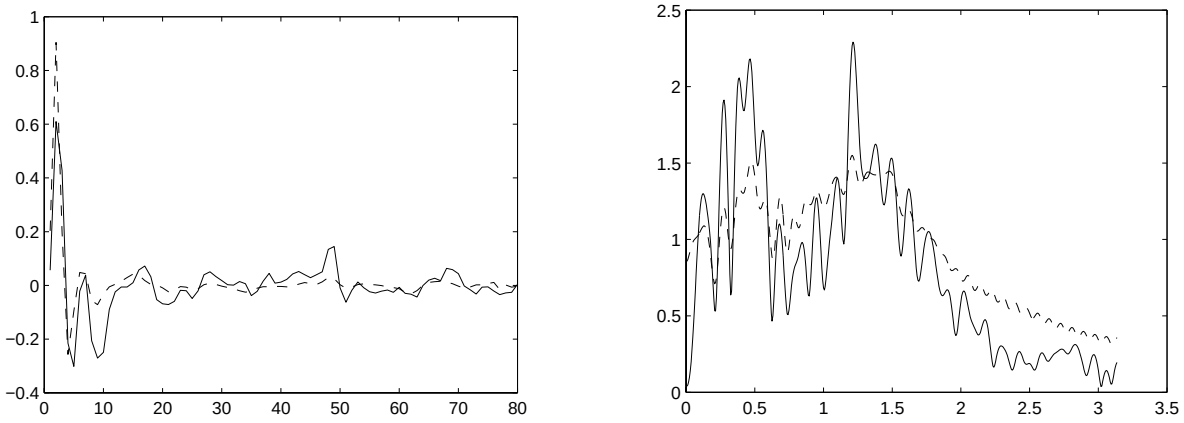

Fig. 2. Blind identification result using the regularized least-squares (LS) method. Left: the time domain impulse response $h_1(n)$ (n is the x-axis). Right: the frequency domain axis $|H(\omega)|$ (ω is the x-axis). The solid curve is the true channel; the dashed curve is the estimate. The regularization parameters: $2\sigma^2\mu_x = 300$, $2\sigma^2\mu_h = 1500$.

In matrix form, that's equivalent to

$$\left[C(\underline{y}_2)| - C(\underline{y}_1)\right] \left[\begin{array}{c} \underline{h}_1 \\ \underline{h}_2 \end{array} \right] = 0, \qquad (2)$$

where $C(\underline{y}_i)$ is the operator of convolving with $\underline{y}_i$. We use the shorthand notation

$$Y\underline{h} = 0, \qquad (3)$$

where $Y = \left[C(\underline{y}_2)| - C(\underline{y}_1)\right]$ is computed from the observations $\underline{y}$. It has been shown in [1, 2] that under the identifiability conditions discussed in Section 1, Y is rank-deficient by only 1. Therefore, the channel $\underline{h}$ satisfying (3) is the eigenvector (up to a scaling) corresponding to the zero eigenvalue of $Y'Y$. This analysis can also be generalized to more than two channels. For details, see for example [4].

In the presence of noise, cross correlation methods can be unreliable, because the eigenvectors of $Y'Y$ are very sensitive to perturbations in $\underline{y}$. This calls for regularization using prior knowledge about $\underline{h}$. We design a training process to acquire such knowledge. The sound source is placed in K random positions in the microphone array lab. In each position, we measure the channels and obtain $\underline{h}_{expk}$. It is often reasonable to assume that $\underline{h}$ can be well approximated by the linear interpolation of these training vectors. Let $H = (\underline{h}_{exp1}, \underline{h}_{exp2}, \ldots, \underline{h}_{expK})$. Hence, $\underline{h} \in \text{Range}(H)$.

We design the regularization penalty to be the distance from the estimate $\underline{h}$ to the range space: $\Phi_h(\underline{h}) = ||(I - P_H)\underline{h}||^2$, where $P_H = H(H'H)^{-1}H'$ is the projection operator. This is illustrated in Fig. 3.

The cost function to be minimized is

$$\mathcal{E} = ||Y\underline{h}||^2 + \mu_h||(I - P_H)\underline{h}||^2. \qquad (4)$$

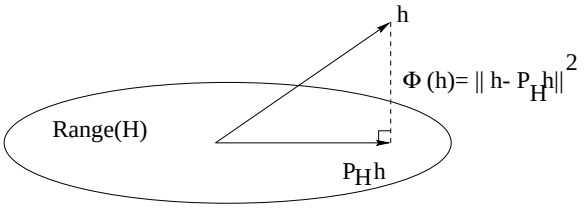

Fig. 3. Projection onto Range(H).

The solution $\underline{h}$ is then the eigenvector corresponding to the smallest eigenvalue of $(Y'Y + \mu_h(I-P_H)'(I-P_H))$. Efficient algorithms can be used to find such a solution, for example [9].

Note that Y is of size $P(P-1)(N-L+1)/2 \times PL$, where N is the number of observed samples. Computing $Y'Y$ directly is not feasible when N is large. Therefore, we compute $Y'Y$ for the first $2L+1$ samples, then update $Y'Y$ for every new incoming M samples, where M can be as small as one. Such an incremental updating strategy can use a forgetting factor to ensure that $Y'Y$ adapts to varying acoustic conditions.

3.1. Experimental Results

We have used the same setting as described in Section 2.2. Fig. 4 plots the identified channel corresponding to the first microphone, after processing 40,000 samples.

As Fig. 4 (a) shows, the estimate (the dashed curve) tracks its true value (the solid curve) very closely, especially in the beginning, where most of the energy resides. The spectral domain plot in Fig. 4 (b) also shows a good fit. The SNR in estimating $|H(\omega)|$ is

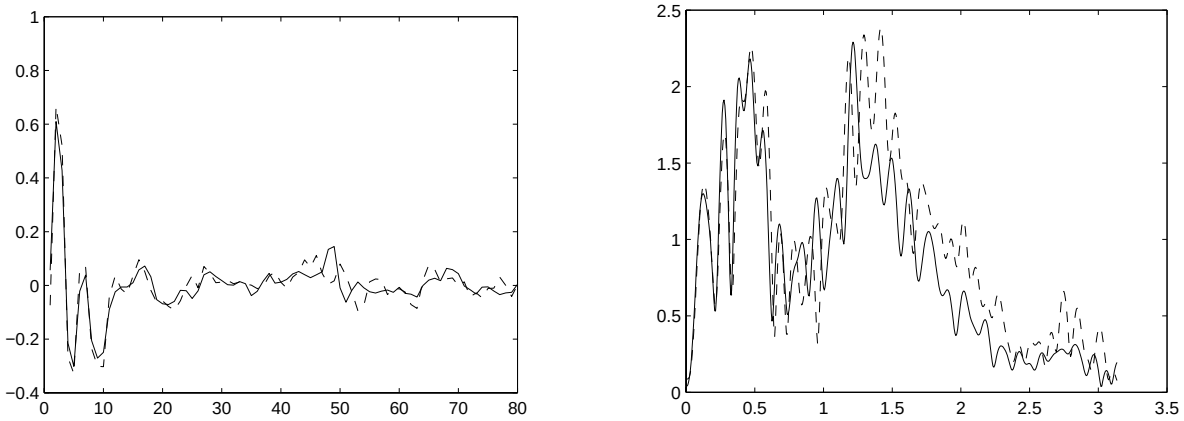

Fig. 4. Blind identification result using the regularized cross correlation (CR) method. Left: the time domain impulse response $h_1(n)$ (n is the x-axis). Right: the frequency domain axis $|H(\omega)|$ (ω is the x-axis). The solid curve is the true channel; the dashed curve is the estimate. The regularization parameter is $2\sigma^2\mu_h = 1$. This plot is drawn after 40,000 samples are processed.

10.9 dB. The restored $\underline{\hat{x}}$ is much closer (than any of the microphone signals) to the clean signal $\underline{x}$ in sound quality, and the SNR gain over the delay-and-sum estimate is 10.1 dB. Those SNR numbers are better than those for the LS method, which is expected since the CR method uses the room response measurements.

4. DISCUSSION

The two regularization approaches, presented in Sections 2 and 3, should be used for different applications, based on the available prior knowledge. The regularized LS method uses weak assumptions about both the signal and the channels to help estimation, and the two regularization terms balances each other. It does not require any specific training of the channel impulse responses. The optimization algorithm gradually learns the underlying channels. The method is recommended when the sound signal is mostly speech, and the room environment is likely to vary (e.g., in a conference room with many people, or for a handheld device).

The regularized CR method in Section 3, needs training to get $H = (\underline{h}_{exp1}, \underline{h}_{exp2}, \ldots, \underline{h}_{expK})$, and thus it is sensitive to large variations in the acoustics of the room. On the other hand, it does not rely on signal statistics, and is applicable to nonspeech signals such as music. This method is recommended in situations where training is meaningful, e.g. in many office environments.

Acknowledgment: the authors are thankful to Dr. Dinei Florêncio, Dr. Alex Acero, Dr. Patrice Simard, Prof. Yingbo Hua, and Bradford Gillespie for cooperation and discussions.

5. REFERENCES

[1] G. Xu, H. Liu, L. Tong, and T. Kailath, "A least–squares approach to blind channel identification," *IEEE Trans. on Signal Processing*, vol. 43, no. 12, pp. 2982–2993, Dec. 1995.

[2] Y. Hua and M. Wax, "Strict identifiability of mulitple FIR channels driven by an unknown arbitrary sequence," *IEEE Trans. on Signal Processing*, vol. 44, no. 3, pp. 756–759, Mar. 1996.

[3] M. I. Gurelli and C. L. Nikias, "EVAM: An eigenvector–based algorithm for multichannel blind deconvolution of input colored signals," *IEEE Trans. on Signal Processing*, vol. 43, no. 1, pp. 134–149, Jan. 1995.

[4] Y. Hua, "Fast maximum likelihood for blind identification of multiple FIR channels," *IEEE Trans. on Signal Processing*, vol. 44, no. 3, pp. 661–672, Mar. 1996.

[5] G. Harikumar and Y. Bresler, "Analysis and comparative evaluation of techniques for multichannel blind equalization," in *Proc. 8th IEEE Signal Processsing Workshop Statistical and Array Signal Processing*, Corfu, Greece, June 1996, pp. 332–335.

[6] D. T. M. Slock and C. B. Papadias, "Further results on blind identification and equalization of multiple FIR channels," in *Proceedings of ICASSP*, Detroit, MI, 1995, pp. 1964–1967.

[7] D. G. Luenberger, *Linear and Nonlinear Programming*, New York, NY: Addison Wesley, 1984, Chap. 7.

[8] James M. Kates, "Signal processing for hearing aids," in *Applications of Digital Signal Processing to Audio and Acoustics*, Mark Kahrs and Karlheinz Brandenburg, Eds., chapter 6. Kluwer, 1990.

[9] K. Abed-Meraim, S. Attallah, A. Chkeif, and Y. Hua, "Orthogonal Oja algorithm," *IEEE Signal Processing Letters*, vol. 7, no. 5, pp. 116–119, May 2000.