MULTI-CHANNEL ECHO CONTROL BY MODEL LEARNING

Majid Fozunbal, Ton Kalker, and Ronald W. Schafer

Hewlett-Packard Laboratories
Palo Alto, CA 94304, USA
majid.fozunbal@hp.com

ABSTRACT

This paper describes a multichannel echo control algorithm that learns a model of the room while tracking its changes. Using the history of echo path estimates, the algorithm gradually infers an eccentric ellipsoid as the space for echo paths (impulse responses). The directions of skewness of the ellipsoid form a basis for a low dimensional affine space (linear manifold) containing the principle components of echo paths. Assigning a high priority to the principle components, the algorithm reduces the dimension of the search space to combat the non-uniqueness problem. We implemented the algorithm on a real-time software platform running on a Xeon 3.4 GHz processor at a sampling rate of 44.1 KHz. In tested practical setups, once the model was mature, the algorithm demonstrated high stability and accuracy without the need to uncorrelate the excitation signals. For a 3×3 multichannel setup, it takes less than 4 seconds to reduce echo by about 22 dB in a regular conference room.

Index Terms— Echo control, learning, manifold, non-uniqueness, and principle components.

1. INTRODUCTION

Recently there has been increasing market demand for immersive telepresence systems. Such systems promote collaboration among remote participants by creating a compelling illusion of being in the same place. Enabling spatial audio in such systems has been a big challenge since it requires a high fidelity multichannel echo control. Following the pioneering work of Sondhi et al. [1], the problem has become an active area of research. A major challenge in these systems is the correlation of the excitation signals sent to the loudspeakers. This phenomenon manifests itself as an ill-conditioned search for echo path estimates, a problem often called the *non-uniqueness* problem resulting in unstable control algorithms [2].

A variety of approaches have been proposed to combat non-uniqueness [2], [3]. A majority of them uncorrelate the excitation signals via nonlinear or time-variant operators. These algorithms are robust and converge quickly. The drawback of these methods is the distortion of spatial and temporal attributes of the audio signals. Such distortion impairs the user experience in telepresence systems. Another category of algorithms constrains the search for the unknown parameters. Some update a portion of parameters, in the time domain or frequency domain, in each new estimation [4]. Some others approximate the space of echo paths by a finite number of set-theoretic constraints [5], [6]. These algorithms do not distort the excitation signals. They regularize the convergence by applying generic constraints on the search space. Thus, they often result in a tradeoff between stability and accuracy, i.e., level of the residual echo.

Can we develop an algorithm that does not distort the excitation signals and converges quickly and accurately? An insightful approach to answer this question is through the principle of *description complexity*[1] [7]. Suppose $\boldsymbol{h}$ represents the underlying echo paths between the loudspeakers and a microphone in a room. For example, Figure 1, depicts three impulse responses from three loudspeakers to a microphone in a typical medium-size conference room. Let $\boldsymbol{h}$ be the vector concatenation of these impulse responses. Suppose $\mathcal{M}$ is a *model* for the space of echo paths. Intuitively, a model could be any regularity that is common among the echo paths. For example, it can be seen in Figure 1 that all the depicted impulse responses decay after a 50 ms period. They all have similar timings between the first direct arrivals and the the first reflections. Using the principle of *two-part code* [7], for a model, $\mathcal{M}$, the complexity of description of $\boldsymbol{h}$, or intuitively the information required to specify $\boldsymbol{h}$, is described by

$$C(\mathcal{M}) + C(\boldsymbol{h}|\mathcal{M}). \tag{1}$$

Here, $C(\mathcal{M})$ denotes the description complexity of $\mathcal{M}$ and $C(\boldsymbol{h}|\mathcal{M})$ is the description complexity of $\boldsymbol{h}$ when $\mathcal{M}$ is known. If the echo control knows $\mathcal{M}$, it needs the additional information $C(\boldsymbol{h}|\mathcal{M})$ to identify $\boldsymbol{h}$. Thus, if $\mathcal{M}$ is such that $C(\boldsymbol{h}|\mathcal{M})$ is sufficiently small, then $\boldsymbol{h}$ can be identified without uncorrelating the excitation signals. In this regard, two main questions are: 1) Does such a model exist? 2) How can an algorithm learn it?

[1] A precise definition of description complexity is beyond the scope of this paper. Here, the notion is used to provide a high level intuition on model learning.

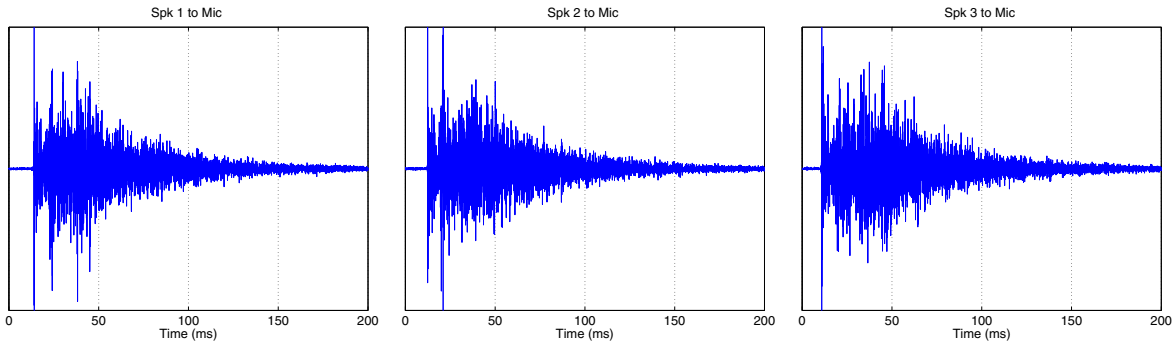

Fig. 1. Impulse responses from (a) Loudspeaker 1, (b) Loudspeaker 2, (c) Loudspeaker 3 to microphone. The results are obtained by exciting the system using independent white Gaussian noise sources.

While we do not have an answer for the first question, we attempt to answer the latter one. We introduce an algorithm that tackles the non-uniqueness problem by learning a model of the room. A key attribute of the non-uniqueness problem is the large number of unknown parameters. These parameters, however, are not necessarily independent. Recently, methods have been proposed to extract a dependence model among parameters and reduce the number of independent unknown parameters. Such methods have been studied under the umbrella of *manifold* modeling [8]. In simple terms, a manifold is a space in which every point has an open neighborhood which resembles a a Euclidean space, with a dimension no larger than the dimension of the space wherein its global structure lies. The surface of a sphere or a donut are some examples.

Our algorithm is inspired by principle component analysis, the simplest, yet most common linear manifold learning [8]. The idea is to pursue a basis such that a major portion of the energy of the impulse responses lies in a small number of basis vectors. For practical reasons, the algorithm utilizes the idea by an evolving eccentric ellipsoid that is skewed along with the principle components. The algorithm learns this ellipsoid, dynamically. Upon initialization, it is blind to the room environment where the model is a ball (or it may have some preprocessing data to start from an ellipsoid model). As it interacts with the room, it uses the obtained *qualified* estimates to gradually transform the ball to an eccentric ellipsoid.

Qualified estimates satisfy two criteria: 1) they correspond to a new echo path, and 2) they are estimates of the echo paths with high accuracy and high confidence. The latter is satisfied by uncorrelating the excitation signals during the learning phase. Using the qualified estimates, the algorithm dynamically improves the inferred model till it becomes mature. Once the model is mature, measured by the number of qualified estimates used in inference, the algorithm saves it for the future use. After this period, the algorithm does not uncorrelate the excitation signals and relies on the inferred model to track the changes in the room.

2. SETUP

Assume a sampling frequency of 44.1 KHz and let n denote the discrete time index. Assume there are M loudspeakers and M microphones in a room. For the sake of clarity, we may consider an arbitrary microphone and conduct all the derivations for this microphone. Hence, throughout the paper, we drop indexing of microphones. At any time instance n, the *room echo path*, i.e., acoustic coupling between loudspeaker i and the microphone is characterized by a vector $h_{i,n} \in \mathbb{R}^L$. In a typical conference room with a *reverberation time* of 200 ms, the size of the echo path impulse response is roughly $L \approx 9000$ taps. Figure 1 depicts three impulse responses from three loudspeakers to a microphone in a typical medium-size conference room.

For the sake of conciseness in derivations, we define a concatenated vector

$$\boldsymbol{h}_n = [h_{1,n}, h_{2,n}, \ldots, h_{M,n}]$$

to denote the overall impulse response of the room. Let $x_i(n)$ denote the audio signal played through loudspeaker i. We define the augmented vector

$$\begin{aligned}\boldsymbol{x}(n) = [&x_1(n), \ldots, x_1(n-L+1), \ldots, \\ &x_M(n), \ldots, x_M(n-L+1)]\end{aligned}$$

to denote the overall *excitation* signal of the system. Thus, the recorded signal at the microphone is described by

$$y(n) = s(n) + \boldsymbol{h}_n' \boldsymbol{x}(n). \tag{2}$$

Here, the signal $s(n)$ represents all locally generated audio signals, i.e., people, audio devices, or noise sources. The second term, $\boldsymbol{h}_n' \boldsymbol{x}(n)$, represents the *multichannel echo*. As people and objects in the room move, the acoustic coupling between loudspeakers and microphone changes. To control the echo, the system needs to continuously track these changes. For this purpose, the system works in a block-wise manner. That is, it takes a new action, $\hat{\boldsymbol{h}}_m$ every N_d ("d" refers to

decision) time samples. The decision times are called *decision epochs* and indices $m = 1, 2, \cdots$ are used to label them. Correspondingly, the duration between two consecutive decisions, N_d, is called a *decision period*. Note that each decision remains the same for any n such that $\lfloor n/N_d \rfloor = m$.

Assume there is no double-talk, i.e., $s(n) \approx 0$. Then,

$$f_m(\hat{\boldsymbol{h}}_m) = \sum_{k=(m-q)N_d+1}^{mN_d} |y(k) - \hat{\boldsymbol{h}}_m' \boldsymbol{x}(k)|^2 \qquad (3)$$

measures the amount of error incurred by deciding to use the estimate $\hat{\boldsymbol{h}}_m$. Here, q denotes the amount of *overlap size* in blocks. There are several factors contributing to how to choose N_d and q: computational resources, sampling frequency, estimation confidence, non-stationarity of speech signals, and movements in the room are all important factors. In practice, a decision period of about 5 ms to 20 ms and an overlap size of 2 to 4 times the decision period is common.

3. THE ALGORITHM

To find the best estimate for the impulse response, the algorithm seeks to find an estimate $\hat{\boldsymbol{h}}_m$ that minimizes $f_m(\hat{\boldsymbol{h}}_m)$. Because of the correlation among excitation signals, however, the algorithm will need to solve a system of normal equations that is either under-determined or severely ill-conditioned.

To combat this problem, we need to find a model $\mathcal{M}$ to limit the search space. In this regard, principle component analysis (PCA) is one of the simplest yet most common techniques. Suppose we have a finite set $\{\boldsymbol{h}_\theta\}_{\theta\in\Theta}$ of sample impulse responses of the room. PCA treats the given set as independent random samples. It computes the empirical average and covariance matrix of the set

$$\bar{\boldsymbol{h}} = \frac{1}{|\Theta|}\sum_\theta \boldsymbol{h}_\theta, \ \Lambda = \frac{1}{|\Theta|}\sum_\theta (\boldsymbol{h}_\theta - \bar{\boldsymbol{h}})(\boldsymbol{h}_\theta - \bar{\boldsymbol{h}})'$$

as an estimate for the mean and covariance of the underlying probability density function. The pair of $(\bar{\boldsymbol{h}}, \Lambda)$ forms the model $\mathcal{M}$. The large eigenvectors of Λ determine the principle directions in which the model is stretched. By keeping the directions that the model is stretched and truncating the directions in which it is condensed the data is compressed. Let

$$\Lambda = [U\ V] \begin{bmatrix} \Sigma & 0 \\ 0 & \Delta \end{bmatrix} \begin{bmatrix} U' \\ V' \end{bmatrix}$$

denote the singular value decomposition of Λ where Σ is a $d \times d$ diagonal matrix containing the d largest eigenvalues. The columns of U define a basis for a d-dimensional subspace that best approximates $\{\boldsymbol{h}_\theta - \bar{\boldsymbol{h}}\}_{\theta\in\Theta}$. In other words,

$$\Lambda_{(d)} = U\Sigma U'$$

denote the best d-dimensional approximation for Λ in a minimum mean-square sense. Here, V is the orthogonormal complement of U. That is $V'U = 0$. Intuitively, if $\{\boldsymbol{h}_\theta\}$ is a good representative for all possible impulse responses, we can expect with high probability that for any impulse response $\boldsymbol{h}$

$$V'(\boldsymbol{h} - \bar{\boldsymbol{h}}) \approx 0. \qquad (4)$$

Equation (4) defines a affine space (linear manifold) in $\mathbb{R}^{ML}$ which serves as a model for the space of impulse responses. Using this model, now we can try to solve for

$$\begin{aligned} \min_{\boldsymbol{h}} \quad & f_m(\boldsymbol{h}) \\ \text{subject to} \quad & V'(\boldsymbol{h} - \bar{\boldsymbol{h}}) \approx 0. \end{aligned} \qquad (5)$$

In this optimization, we have an additional number of $ML - d$ (the rank of V) equations that help to combat the non-uniqueness problem. If $d \le L$, then the non-uniqueness problem is completely resolved. In practice, we replace (5) by the ellipsoid constraint

$$(\boldsymbol{h} - \bar{\boldsymbol{h}})'\Lambda^{-1}(\boldsymbol{h} - \bar{\boldsymbol{h}}) \le 1 \qquad (6)$$

that enforces (5) and limits the energy of estimates.

Table 1. The process of learning the ellipsoid model.

1. Initialization: $\beta_0^{-1} = ML$, $\bar{\boldsymbol{h}}_0 = 0$
   $\Psi_0^{-1} = \beta_0 \mathcal{E} I$, $\Lambda_0^{-1} = \mathcal{E} I$,
2. Recursion: At decision epoch m:
   - if $\hat{h}_{m-1}$ is not qualified, then
     $\beta_m^{-1} = \beta_{m-1}^{-1}$, $\bar{\boldsymbol{h}}_m = \bar{\boldsymbol{h}}_{m-1}$,
     $\Lambda_m^{-1} = \Lambda_{m-1}^{-1}$, $\Psi_m^{-1} = \Psi_{m-1}^{-1}$
   - else
     $\beta_m^{-1} = \beta_{m-1}^{-1} + 1$, $\bar{\boldsymbol{h}}_m = (1-\beta_m)\bar{\boldsymbol{h}}_{m-1} + \beta_m \hat{\boldsymbol{h}}_{m-1}$
     $\Psi_m^{-1} = \Psi_{m-1}^{-1} - \dfrac{\Psi_{m-1}^{-1}\hat{\boldsymbol{h}}_{m-1}\hat{\boldsymbol{h}}_{m-1}'\Psi_{m-1}^{-1}}{1 + \hat{\boldsymbol{h}}_{m-1}'\Psi_{m-1}^{-1}\hat{\boldsymbol{h}}_{m-1}}$
     $\Lambda_m^{-1} = \beta_m^{-1}\big(\Psi_m^{-1} - \dfrac{\Psi_m^{-1}\bar{\boldsymbol{h}}_m\bar{\boldsymbol{h}}_m'\Psi_m^{-1}}{\bar{\boldsymbol{h}}_m'\Psi_m^{-1}\bar{\boldsymbol{h}}_m - \beta_m}\big)$
3. Termination: if $\beta_m^{-1} = \alpha ML$ for some $\alpha > 1$
   - set $\bar{\boldsymbol{h}} = \bar{\boldsymbol{h}}_m$ and $\Lambda = \Lambda_m$.
   - stop uncorrelation.

3.1. Model learning

The first time the system is put to use $\{\boldsymbol{h}_\theta\}_{\theta\in\Theta}$ is empty. Hence, initially the ellipsoid constraint is the Euclidean unit ball constraint scaled by an energy factor. More precisely, at decision epoch 0, the constraint is specified by $\bar{\boldsymbol{h}}_0 = 0$ and $\Lambda_0 = \frac{1}{\mathcal{E}} I$ where $\mathcal{E}$ represents the maximum energy of the room. Gradually, as time passes, the system obtains qualified estimated impulses responses. Using these estimates, the algorithm learns the model through a process that is summarized in Table 1.

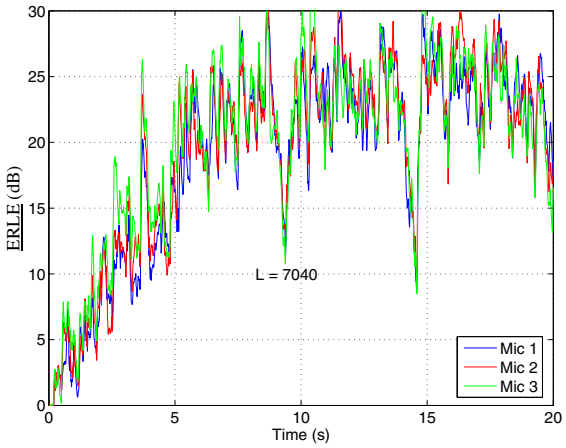

Fig. 2. ERLE after the maturity of the model. The decision period is equivalent to 12 ms.

3.2. Estimating the echo paths

This ellipsoid, specified by the process in Table 1, serves as a trust region in the search for the next best estimate. Suppose the previous estimate is $\hat{\boldsymbol{h}}_{m-1}$ and the system is not in a double-talk situation. To find the next estimate, the algorithms seeks a direction p such that

$$\min_{p} \quad f_m(p + \hat{\boldsymbol{h}}_{m-1})$$
$$\text{subject to } \|p + \hat{\boldsymbol{h}}_{m-1} - \bar{\boldsymbol{h}}_m\|^2_{\Lambda_m^{-1}} \leq 1$$

Solving this problem for p, we obtain

$$(\nabla^2 f_m(\hat{\boldsymbol{h}}_{m-1}) + \lambda\Lambda_m^{-1})p_m = -\nabla f_m(\hat{\boldsymbol{h}}_{m-1}) - \lambda\Lambda_m^{-1}(\hat{\boldsymbol{h}}_{m-1} - \bar{\boldsymbol{h}}_m)$$

where λ is the Lagrange multiplier. To be robust, the algorithm takes $\mu_m \ll 1$ as the learning factor and computes

$$\hat{\boldsymbol{h}}_m = \hat{\boldsymbol{h}}_{m-1} + \mu_m p_m \tag{7}$$

as the estimate for decision epoch m. In the current development, the algorithm has $\mu_m = O(\frac{1}{m-m_o})$ where m_o denotes the decision epoch corresponding to the last major change in the impulse response.

4. NUMERICAL RESULTS

To demonstrate the convergence performance of the algorithm, we use the commonly used criterion *echo-return loss enhancement* (ERLE). Figure 2 plots the ERLE (lower bound) for all three microphones in a 3×3 audio setup. The excitation signals are spatial recordings of a single speech source. The excitation signals are played through the three loudspeakers with no uncorrelation. The results are shown after the model matures for a filter length of 7040 taps. The algorithm runs real-time on a software platform taking 60% of a Xeon 3.4 GHz. The floor levels and high variation seen in the plots are due to the background noise and speech signal variations.

5. CONCLUSION

We presented a multichannel echo control algorithm that combats the non-uniqueness problem by learning a model of the room. The model is a low dimensional linear manifold that contains the principle components of the echo paths. In tested practical setups, the algorithm demonstrated high stability and accuracy without uncorrelating the excitation signals, hence, preserving the audio quality.

Drafting this paper, we had two questions in mind: 1) Does there exist a model that can completely solve the non-uniqueness problem? 2) How can such a model be learned? We did not answer the first question and it remains open to future research. However, we attempted to answer the latter question by introducing a methodology of model learning in echo control. In this regard, learning the model via principle component analysis (PCA) was chosen due to its simplicity to implement and to demonstrate the concept. While the observed performance demonstrates high stability and accuracy, this does not necessarily mean the model solves the non-uniqueness problem in its pure mathematical terms. A better model for the space of echo paths may be a nonlinear manifold that should be approximated with the more sophisticated techniques of manifold learning.

References

[1] M. Sondhi, D. Morgan, and J. L. Hall, "Stereophonic acoustic echo cancellation-an overview of the fundamental problem," *IEEE Signal Processing Letters*, vol. 2, no. 8, pp. 148–151, 1995.

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

[3] M. Ali, "Stereophonic acoustic echo cancellation system using time-varying all-pass filtering for signal decorrelation," *Proc. IEEE Intern. Conf. Acoustics, Speech, and Signal Proc.*, pp. 3689–3692, 1998.

[4] T. Gaensler and J. Benesty, "Multichannel acoustic echo cancellation: whats new?" *Proc. 7th IEEE International Workshop on Acoustic Echo and Noise Control*, 2001.

[5] M. Fozunbal, "On regularization of least square problems via quadratic constraints," *Proc. IEEE ICASSP*, Apr. 2007.

[6] M. Yukawa, N. Murakoshi, and I. Yamada, "Efficient fast stereo acoustic echo cancellation based on pairwise optimal weight realization technique," *EURASIP Journal on Applied Signal Processing*, pp. 1–15, 2006.

[7] P. D. Grunwald, *The Minimum Description Length Principle*. The MIT Press, 2007.

[8] A. Gorban, B. Kegl, D. Wunsch, and A. Zinovyev, *Principal Manifolds for Data Visualisation and Dimension Reduction*. Berlin: Springer, 2007.