A STATISTICAL PHYSICS PERSPECTIVE OF ACOUSTIC ECHO CONTROL

Peter Riegler

ABB Corporate Research
Wallstadter Str. 59, D-68526 Ladenburg
peter.riegler@de.abb.com

ABSTRACT

In the past years, statistical physics has created methods to derive algorithms which are optimal, *i.e.* these algorithms maximize the convergence speed. Most importantly, these algorithms can be calculated starting from first principles. This paper shows how to derive such optimal algorithms for acoustic echo cancellation using statistical physics. The results are compared to standard algorithms of the field.

1. INTRODUCTION

Statistical physics deals with systems of many interacting degrees of freedom. Traditionally, it investigates many-particle systems such as gases or solids. The methods developed have been successfully expanded to other "many-particle" systems such as neural networks, traffic or markets during the past decade. Some of these methods have been designed for the on-line estimation of parameters and, hence, can directly be applied to acoustic echo control (AEC).

The most distinct feature of the statistical physics approach is that algorithms can be *calculated* from first principles. This is done by varying the update rule with respect to the algorithm. By construction the resulting algorithms are optimal in the typical case. Moreover, these algorithms show desirable features, *e.g.* automatic stepsize control, without having to put them in explicitly.

2. GENERAL ALGORITHM

In this work I will restrict to the class of algorithms where the change of the ith filter coefficient is proportional to the ith component of the signal vector. A purely sequential update rule for the filter coefficients $\underline{c}(n)$ at time sample n can then be written in general form as

$$\underline{c}(n+1) = \underline{c}(n) + F\left(\underline{c}(n), \underline{x}(n), y(n)\right) \underline{x}(n) \qquad (1)$$

where $\underline{x}(n)$ is the vector that comprises the N latest far-end signal samples and $y(n)$ is the local signal at time n. The local signal is given by $y(n) = \underline{g}^\top \underline{x}(n) + \xi(n)$, with $\underline{g}$ being the room impulse response and ξ any additional noise or double talk. In this paper I will assume, that $||\underline{g}||$ is known or reliably estimated from the data $y(n)$. Without loss of generality $||\underline{g}||$ can then chosen to be unity.

Further assuming that there is no *a priori* knowledge about the structure of $\underline{g}$, the weight function F in (1) can only depend on the scalar products of $\underline{c}$ and $\underline{x}$ due to symmetry. Consequently, (1) can be written as

$$\underline{c}(n+1) = \underline{c}(n) + \frac{1}{N} f\left(h_c, c, y\right) \underline{x}(n), \qquad (2)$$

where $h_c = \underline{c}^\top \underline{x}/c$ and $c = ||\underline{c}|| = \sqrt{\underline{c}^\top \underline{c}}$ at time sample n. Note that $||\underline{x}||^2 = \mathcal{O}(N)$. The algorithm used for updating the filter coefficients is then completely characterized by the particular realization of the weight function f in (2).

3. VARIATIONAL APPROACH

The statistical physics approach has two main ingredients [1, 2]: the treatment of the update rule in the so-called *thermodynamic limit* $N \to \infty$ and the determination of the optimal algorithm by a variation with respect to f in (2). An alternative, but equivalent approach based on Bayesian statistics is outlined in [5].

For deriving optimal algorithms it turns out that instead of using $e = y - \underline{c}^\top \underline{x}$ the overlap $\rho = \underline{g}^\top \underline{c}/c$ is a more convenient error measure. Clearly, as $\underline{c} \to \underline{g}$ one will have $\rho \to 1$.

Multiplying (2) with $\underline{g}$ and $\underline{c}$, respectively, one obtains after normalization:

$$\begin{aligned} \rho(n+1) &= \rho(n) + \frac{1}{N}\left(f(h_g - \rho(n) h_c) - \frac{\rho(n)}{2} f^2 \right) \\ c(n+1) &= c(n) + \frac{c}{N}\left(2 f h_c + f^2\right). \end{aligned} \qquad (3)$$

Here and in the following the arguments of f are omitted for simplicity.

Taking the thermodynamic limit, *i.e.* for large N, the stochastic difference equations (3) can be written as a set of differential equations for ρ and c:

$$\frac{d\rho}{d\tau} = \left\langle f(h_g - \rho h_c) - \frac{\rho}{2} f^2 \right\rangle_{P(h_c, h_g, y)} \qquad (4)$$

$$\frac{dc}{d\tau} = c\left\langle 2fh_c + f^2\right\rangle_{P(h_c,h_g,y)}, \quad (5)$$

where $\tau = n/N$ defines the relevant time scale and the average $\langle\cdot\rangle$ is over the joint distribution of y, h_c, and $h_g = \underline{g}^\top \underline{x}$. It needs to be stressed that ρ and c do not fluctuate for $N \to \infty$, since they are self-averaging in this limit, *i.e.* their distribution becomes sharply peaked (basically ρ and c take the role of mean and variance of $\underline{c}$, respectively).

The set of differential equations (4,5) can be solved numerically for any given algorithm f. Thus, one obtains the convergence properties in the typical case. Moreover, the asymptotic behavior for $\tau \to \infty$ can often be obtained analytically.

Eq. (4) determines the time evolution of the overlap ρ, which is a measure of the performance of the algorithm f. Clearly, the optimal algorithm f_{opt} is characterized by the fastest change $d\rho/d\tau$ of the performance ρ towards 1. Thus, f_{opt} can be determined by varying $d\rho/d\tau$ w.r.t. the weight function f

$$\frac{\delta}{\delta f}\left(\frac{d\rho}{d\tau}[f]\right) = 0, \quad (6)$$

resulting in the optimal algorithm

$$f_{opt} = \frac{1}{\rho}\left\langle h_g\right\rangle_{P(h_g|h_c,y)} - h_c \quad (7)$$

with an average over h_g conditional on the (accessible) quantities h_c, y. By construction f_{opt} is that algorithm that converges as fast as possible.

One might object that (7) is not a realistic algorithm as it explicitly depends on the performance ρ, *i.e.* the actual error, which is not accessible, in general. However, inserting (7) into the differential equations (4,5) one obtains

$$\frac{d\rho}{d\tau} = \frac{\rho}{2}\left\langle f_{opt}^2\right\rangle_{P(h_c,h_g,y)} \quad (8)$$

$$\frac{dc}{d\tau} = \frac{c}{2}\left\langle f_{opt}^2\right\rangle_{P(h_c,h_g,y)} \quad (9)$$

showing that $\rho(\tau) = c(\tau)$ holds as a general property of f_{opt} if one starts with filter coefficients set to zero in the beginning. Hence, the performance measure always equals the length of the signal vector and both converge to 1 from below. Consequently, ρ can always be replaced by c in (7).

Despite its simple functional form, eq. (7) can be tricky to evaluate for real signals. Nevertheless, it shows which information is essential, namely the conditional distribution of $h_g = \underline{g}^\top \underline{x}$. Thus, any knowledge about the structure of the room impulse $\underline{g}$ and the distribution of far-end signals $\underline{x}$ strongly influences the performance of the optimal filter algorithm (7). Any reduction in the uncertainty of the model about these quantities will pay back via an improvement of the algorithm. Thus, the algorithm (7) is optimal *given* the knowledge one has about the distribution of h_g.

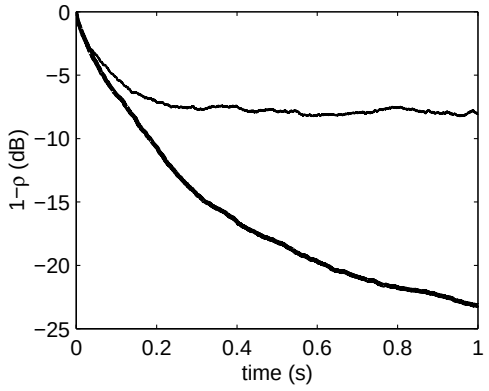

Figure 1: Comparison of the performance of the algorithms (11) (thick line) and (10) (thin line) for white noise as a far-end signal. Algorithm (10) does not converge due to the presence of a local distortion with signal to noise ratio of 6.5dB. Here and in the following figures the sampling rate has been chosen to be 10kHz.

A note is in place here regarding the underlying assumption that there is no *a priori* knowledge about the structure of room impulse response $\underline{g}$. Relaxing this assumption does not have any impact on the feasibility of the method described in this section. If one actually has knowledge about $\underline{g}$ in a parameterized form, the variational approach can still be applied.

4. EXAMPLES

It is interesting to see that for simple models of the far-end signal and the local distortion f_{opt} reduces to algorithms which are very similar to those being discussed in the literature of AEC.

4.1. Distortion free case

Modeling the far-end signal as white noise and assuming no local distortion being present, one obtains after carrying out the average in (7):

$$\frac{1}{N} f_{opt}\underline{x} = \frac{1}{N}\left(y - h_c\right)\underline{x} \simeq \frac{e(n)\underline{x}(n)}{||\underline{x}(n)||^2}. \quad (10)$$

This basically is the well-known NLMS algorithm, leading to $1-\rho = \exp(-\tau)$, *i.e.* exponentially fast convergence, asymptotically [3].

4.2. Distortion modeled as Gaussian noise

Modeling far-end signal and local distortion as mutually uncorrelated noise results in the optimal algorithm

$$f_{opt} = \alpha\left(\frac{y}{\rho} - h_c\right), \quad \alpha = \frac{1-\rho^2}{1-\rho^2+\sigma^2}, \quad (11)$$

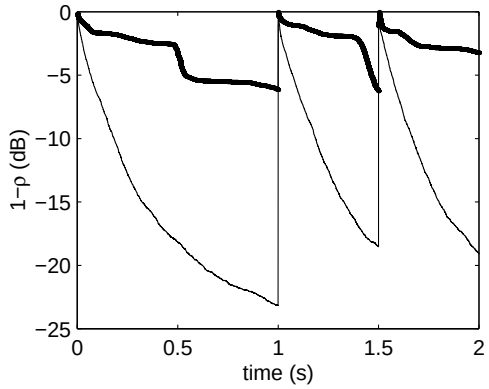

Figure 2: Performance of the algorithm (11) for a varying impulse response which changes abruptly at 1.0s and 1.5s. Depicted are results for white noise (thin line) and speech (thick line). Note that for the latter case (11) is strictly speaking not the optimal algorithm as (11) was computed with white noise as a far-end signal.

where σ^2 denotes the variance of the local distortion. For $\tau \to \infty$ this algorithm convergences like $1 - \rho \propto \tau^{-1/2}$.

As can be seen the necessary stepsize α comes in automatically in (11), exclusively determined by the current performance of the filter and the data. As emphasized in Section 3, the explicit dependence of the stepsize α on the (inaccessible) performance ρ can be removed by replacing ρ by c, since $\rho = c$ for the optimal algorithm. Most remarkably, however, is the explicit deviation of the error term $(y/\rho - h_c)$ of the algorithm (11) from the LMS form $(y - h_c)$ for $\rho \neq 1$. This clearly indicates that LMS error term is only asymptotically optimal, *i.e.* for $\rho \simeq 1$.

Fig. 1 shows the performance of this algorithm compared with that of algorithm (10).

4.3. Varying impulse response

In a real environment the room impulse response is not constant over time. Fig. 2 shows an extreme case where the impulse response changes abruptly several times. As can be seen algorithm (11) perfectly copes with the situation although no *ad hoc* stepsize control has been build in.

The algorithm automatically detects the change of $\underline{g}$ via its dependence of the stepsize on ρ in (11). While $\rho \simeq 1$ just before the change of $\underline{g}$, the overlap $\rho \propto \underline{g}^\top \underline{c}$ becomes 0 just after an abrupt change of the impulse response since then $\underline{g} \perp \underline{c}$ with high probability.

4.4. Correlated far-end signal

Modeling the far-end signal as white noise surely is a crude simplification. In principle one can easily relax this simplification by computing the optimal algorithm according to (7). However, one would need to have a fair idea about the functional form of the distribution of the far-end signals $\underline{x}(n)$ in order to carry out the average over $P(h_g|h_c, y)$. In other words if one knew the distribution of the far-end signal one could readily compute the truly optimal algorithm via (7).

This shows the real problem of construction algorithms for AEC: Each member of the family of optimal algorithms (7) is characterized by the distribution $P(h_g|h_c, y)$. Hence, the less one's uncertainty about this distribution the better the corresponding algorithm, as already pointed out in Section 3. In addition, one needs to be able to actually carry out the average over $P(h_g|h_c, y)$ in (7). The latter will not be possible in general, at least not in an analytical way. Hence, what one eventually needs to do is to find reasonable approximations to the averages in the differential equations (4,5) and the resulting optimal algorithm (7).

How the required averages in (4,5) can be approximated for correlated far-end signals has been shown in [4] for a simple (non-optimal) choice of f . However, the application of this to the optimal algorithm (7) is still a topic for future research.

Here, I will take a simplified approach, assuming that the correlation of the far-end signal is given by a first-order Markov process with time-varying correlation $C(n)$:

$$\langle \underline{x}(n+1)\underline{x}(n) \rangle = C(n+1). \tag{12}$$

$C(n)$ is estimated at each time step and the update of the filter coefficients is performed via (2) with the replacements

$$\underline{x}(n) \quad \to \quad \underline{x}(n) - \tilde{C}(n)\underline{x}(n-1) \tag{13}$$

$$y(n) \quad \to \quad y(n) - \tilde{C}(n)y(n-1), \tag{14}$$

where $\tilde{C}$ is the estimate of C. Essentially this corresponds to pre-whitening the far-end signal.

Fig. 3 compares the performance of algorithm (11) with and without this pre-whitening procedure applied to real speech signals. As has been discussed above, none of these two algorithms is the true optimal one according to eq. (7). Algorithm (11) combined with the pre-whitening procedure outlined above can be viewed as an lower bound on the performance ρ of the truly optimal algorithm for this setting.

5. COMPLEXITY VS. PERFORMANCE

This paper restricts to updates of the form (1) where the update only depends on the most recent far-end signal. An alternative would be to replace (1) by an update of the form

$$\underline{c}(n+1) = \underline{F}(\underline{c}(n), \{\underline{x}(i), y(i)\}_{i=n-m,..,n}) \tag{15}$$

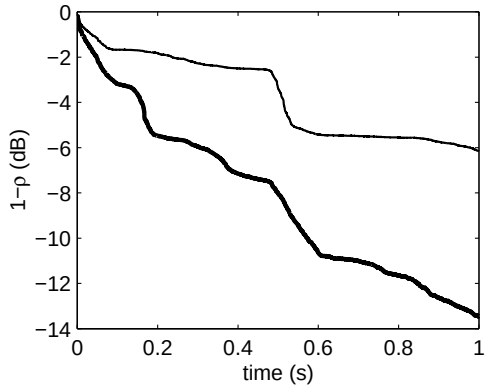

Figure 3: Performance of the algorithm (11) (thin line) and algorithm (11) with additional pre-whitening (thick line). The far-end signal is speech plus local distortion (signal to noise ratio 6.5dB).

depending on the m latest far-end signals. One might hope that such an algorithm leads to faster convergence of the filter coefficients.

Clearly, the algorithm (15) has a higher computational complexity than (1) since it requires the storage and handling of considerably more quantities. Does this higher complexity also cause a better performance (faster convergence)? Opper [5] has recently shown for some neural network applications that the convergence of the optimal algorithm for updates of type (15) is at most faster by a factor of 2 as compared to the optimal algorithm of type (1). While it is currently not clear whether the results of [5] hold for the general case they give some confidence that using the latest far-end signal only is a reasonable simplification of the more general algorithm (15).

6. CONCLUSIONS

The computational methods developed within statistical physics can be used to derive the optimal algorithm for AEC from first principles. Not too surprisingly many features of such optimal algorithms are essentially contained in established algorithms for AEC. However, the computational methods outlined in this paper provide a clear route to improve existing algorithms by showing the essential dependencies of the optimal algorithm.

7. REFERENCES

[1] Kinouchi, O., Caticha, C., J. Phys. **A 25**, 6243, 1992.

[2] Biehl, M., Riegler, P., Stechert, M., Phys. Rev. E **52**, R4624, 1995.

[3] Kinouchi, O., Caticha, C., Phys. Rev. E **52**, 2878, 1995.

[4] Wiegerinck, W., Heskes, T., Europhys. Lett. **28**, 451, 1994.

[5] Opper, M.: *A Bayesian Approach to Online Learning*. In: D. Saad (ed.): *On-Line Learning in Neural Networks*, Cambridge University Press, Cambridge, 1999.