Adaptive Noise Cancellation

P. Comon and V. Zarzoso

January 5, 2010

1 Introduction

In numerous application areas, including biomedical engineering, radar, sonar and digital communications, the goal is to extract a useful signal corrupted by interferences and noises. Noise/interference removal is facilitated when multiple sensors on different locations record the biomedical phenomenon simultaneously.

For instance in recordings taken from the mother's skin during pregnancy, the electrical activity from the fetal heartbeat can be masked by the stronger maternal cardiac activity. In electrocardiogram (ECG) recordings from atrial fibrillation sufferers, the electrical activity from the atria appears mixed with that from the ventricles. In electroencephalogram (EEG) recordings from epileptic patients, epileptic discharges and the brain's background activity contribute simultaneously to the signals measured by scalp electrodes, and can be further corrupted by artifacts such as eye blinks or body movements.

Depending on the modeling that is assumed, fetal ECG extraction, and atrial activity extraction, can be addressed at least in two ways [3]:

- Adaptive noise cancelling (ANC)
- Blind Source Separation (BSS)

The methodology below — classical in biomedical signal processing — is followed:

- analyze the biomedical problem to solve
- choose an appropriate modeling
- propose signal processing techniques, well matched to the model
- evaluate and compare performances on synthetic signals, and then on real data.

2 Theory

Assume that we measure at a sensor output a primary signal $x(t)$ containing a desired signal $d(t)$ contaminated by an interference or a noise $b(t)$. If the contamination is assumed additive, we have:

$$x(t) = d(t) + b(t) \tag{1}$$

Now assume we have at our disposal K sensors recording "reference signals", $z_i(t)$, $1 \leq i \leq K$, correlated with $b(t)$ but uncorrelated with the desired signal $d(t)$. The idea of noise cancellation with references is to filter references $z_i(t)$ so that the filter output approximates $b(t)$, and then to subtract it from the observation.

2.1 Modeling

We use a Finite Impulse response (FIR) of length M, defined by its taps $w_i(k)$, $1 \le k \le M$. After subtraction, the estimation of the desired signal takes the form:

$$\begin{aligned} s(t) &= x(t) - \sum_{i=1}^{K} w_i \star z_i(t) & (2) \\ &= x(t) - \sum_{i=1}^{K} \sum_{k=1}^{M} w_i(k)\, z_i(t-k) & (3) \end{aligned}$$

This relation can be written in a compact way as

$$s(t) = x(t) - \mathbf{w}^{\mathrm{T}} \mathbf{z}(t) \tag{4}$$

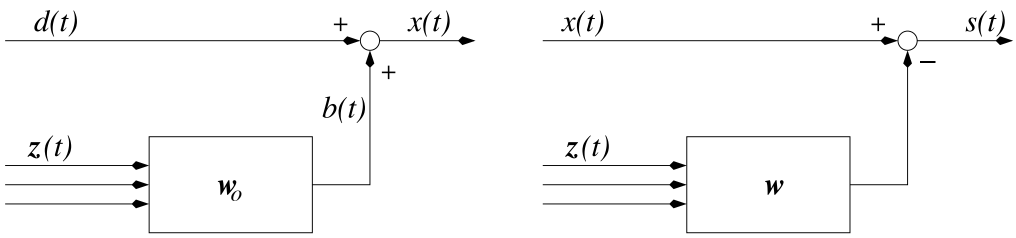
where

$$\begin{aligned} \mathbf{w} &\stackrel{\text{def}}{=} [w_1(0), \ldots, w_1(M), w_2(0), \ldots, w_2(M), \ldots, w_K(0), \ldots, w_K(M)], \\ \mathbf{z}(t) &\stackrel{\text{def}}{=} [z_1(t), \ldots, z_1(t-M), z_2(t), \ldots, z_2(t-M), \ldots, z_K(t), \ldots, z_K(t-M)] \end{aligned}$$

It is important to stress that, by *definition*, $b(t)$ denotes the quantity present in the observation that is linearly related with references $z_i(t)$. Should part of the noise not be represented by the references, it would be put in $d(t)$. In our model, we thus assume that:

$$\exists\, \mathbf{w}_o \text{ such that:} \quad b(t) = \mathbf{w}_o^{\mathrm{T}} \mathbf{z}(t) \tag{5}$$

The influence of a loss in coherence between noise $b(t)$ and reference $\mathbf{z}(t)$ is out of the scope of the present description. The reader may refer to [1] for further reading concerning this modeling error. As a consequence, if filter $\mathbf{w}$ is perfectly estimated, we shall eventually have $d(t) = s(t)$.



2.2 Solution

The optimal filter taps are those minimizing the output power:

$$\mathbf{w}_o = arg\min_w \xi, \quad \xi \stackrel{\text{def}}{=} \mathrm{E}\{s(t)^2\}$$

Because $b(t)$ and the desired signal $d(t)$ are not correlated, this criterion is nothing else but the mean square error:

$$\xi = \mathrm{E}\{(s(t) - d(t))^2\} + constant,$$

which legitimates its choice. It can be shown that the gradient of ξ with respect to $\mathbf{w}$ takes the form

$$\mathbf{g} = -2\ \mathrm{E}\{\mathbf{z}(t) s(t)\}$$

When this gradient is null, we can observe that the optimal filter decorrelates the output $s(t)$ from the reference signal $\mathbf{z}(t)$, which makes sense. This decorrelation may be seen as an orthogonality, in the sense of the scalar product $\langle \mathbf{u}, \mathbf{v} \rangle = \mathrm{E}\{\mathbf{u}^{\mathrm{T}} \mathbf{v}\}$. If signals are Gaussian, it also corresponds to a statistical independence.

2.3 Algorithm

Now, if we plug the definition (4) of $s(t)$ back in the gradient expression, we get the equation below defining the optimal filter $\mathbf{w}_o$:

$$\mathrm{E}\{\mathbf{z}(t)x(t)\} - \mathrm{E}\{\mathbf{z}(t)\,\mathbf{z}(t)^{\mathrm{T}}\}\mathbf{w}_o = 0$$

or, in other words, $\mathbf{w}_o$ is given by the solution of the so-called *normal equation*:

$$\mathbf{w}_o = \mathbf{R}_{zz}^{-1}\mathbf{r}_{zx} \tag{6}$$

if we denote $\mathbf{r}_{zx} \stackrel{\mathrm{def}}{=} \mathrm{E}\{\mathbf{z}(t)\,x(t)\}$ and $\mathbf{R}_{zz} \stackrel{\mathrm{def}}{=} \mathrm{E}\{\mathbf{z}(t)\,\mathbf{z}(t)^{\mathrm{T}}\}$. In practice, the solution of a linear system is required, and may computed with the help of standard routines borrowed from a mathematical library.

Another possibility consists of computing iteratively the solution by minimizing the criterion ξ by a gradient descent:

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} - \mu\,\mathbf{g}(\mathbf{w}^{(t)}) \tag{7}$$

This solution generally requires a much larger cumulated number of operations, but is sometimes preferred because of its smaller peak computational load, and its simplicity of implementation.

In the context of adaptive processing, this algorithm can be simplified one step further, by ignoring the mathematical expectation in the expression (7) of the gradient. In that case, the algorithm is referred to as “stochastic gradient” or ”Least Mean-Square algorithm” (LMS):

$$\mathbf{w}^{(t+1)} = \mathbf{w}^{(t)} + 2\,\mu\,\mathbf{z}(t)\,[x(t) - \mathbf{z}(t)^{\mathrm{T}}\mathbf{w}^{(t)}] \tag{8}$$

The advantage of this algorithm lies in the fact that an estimate of the filter output can be updated at each time step. In addition, the computational load for this update is reduced. Note however that when an acceptable solution has been obtained, the cumulated computational complexity is generally much larger than that required for computing the solution (6) directly (i.e. by a batch processing).

The scalar number μ is called the adaptation step. When the stochastic gradient algorithm is used to cancel interferences with references, one usually talks about *multi-reference adaptive noise cancelling* (MRANC).

2.4 Convergence

The problem with the stochastic gradient algorithm is that parameter μ controls both the step size of the gradient descent, and the implicit exponential averaging of the second order moments. Hence it should be chosen with care. A large value of μ will ease a rather rapid convergence, but will leave a large residual variance about the solution. A small value of μ will lead to a more accurate solution, but will slow down convergence.

In order to study the convergence of the stochastic gradient algorithm, we shall assume that the reference signal is white. In particular, this implies that

$$\mathrm{E}\{\mathbf{z}(t_1)\,\mathbf{z}(t_2)^{\mathrm{T}}\} = \mathbf{R}_{zz}\,\delta(t_1 - t_2)$$

We shall first proceed to a simplification, which does not reduce the generality. Consider the eigenvalue decomposition of matrix $\mathbf{R}_{zz}$ as $\mathbf{R}_{zz} = \mathbf{Q}\,\mathbf{\Lambda}\,\mathbf{Q}^{\mathrm{T}}$, where $\mathbf{Q}$ is orthogonal and $\mathbf{\Lambda}$ is diagonal positive definite. Then one can make the following change of basis:

$$\begin{aligned}\mathbf{z}' &= \mathbf{Q}^{\mathrm{T}}\mathbf{z}\\ \mathbf{w}' &= \mathbf{Q}\,\mathbf{w}\end{aligned}$$

so that the covariance of $\mathbf{z}'$ is now diagonal and equal to $\mathbf{\Lambda}$, whereas we still have $s(t) = x(t) - \mathbf{w}'^{\mathrm{T}}\mathbf{z}'(t)$. Consequently, searching for $\mathbf{w}$ or for $\mathbf{w}'$ are equivalent.

Define the random variable $\mathbf{h}(t) \stackrel{\text{def}}{=} \mathbf{w}'(t) - \mathbf{w}'_o$, where $\mathbf{w}'_o \stackrel{\text{def}}{=} \mathbf{R}_{z'z'}^{-1}\mathbf{r}_{z'x} = \mathbf{Q}\,\mathbf{w}_o$. In this section, we shall study under what conditions $\mathrm{E}\{h(t)\}$ converges to zero. First remark that, from Equation (5), $\mathbf{z}(t)^{\mathrm{T}}\mathbf{w}_o = b(t)$, or equivalently $\mathbf{z}'(t)^{\mathrm{T}}\mathbf{w}'_o = b(t)$. This permits to write

$$x(t) - \mathbf{z}'(t)\mathbf{w}' = x(t) - \mathbf{z}'(t)^{\mathrm{T}}\mathbf{h} - b(t) = d(t) - \mathbf{z}'(t)^{\mathrm{T}}\mathbf{h}.$$

Then from Equation (8), we have

$$\mathbf{h}^{(t+1)} = [\mathbf{I} - 2\,\mu\,\mathbf{z}'(t)\,\mathbf{z}'(t)^{\mathrm{T}}]\,\mathbf{h}^{(t)} + 2\,\mu\,\mathbf{z}'(t)\,d(t) \tag{9}$$

Because $\mathbf{w}'(t)$'s depend only on values of $\mathbf{z}'(k)$ for $k < t$, and because $\mathbf{z}'(t)$ is white, $\mathbf{h}(t)$ is independent of $\mathbf{z}'(t)$. Hence, by taking the mathematical expectation of both sides, one gets:

$$\mathrm{E}\{\mathbf{h}^{(t+1)}\} = [\mathbf{I} - 2\,\mu\;\mathrm{E}\{\mathbf{z}'(t)\,\mathbf{z}'(t)^{\mathrm{T}}\}]\;\mathrm{E}\{\mathbf{h}^{(t)}\} + 2\,\mu\;\mathrm{E}\{\mathbf{z}'(t)\,d(t)\}$$

The last term vanishes, by hypothesis, which allows us to write

$$\mathrm{E}\{\mathbf{h}^{(t+1)}\} = [\mathbf{I} - 2\,\mu\,\mathbf{\Lambda}]^t\,\mathrm{E}\{\mathbf{h}^{(1)}\} \tag{10}$$

This is a geometrical series, which converges if and only if the eigenvalues of matrix $[\mathbf{I} - 2\,\mu\,\mathbf{\Lambda}]$ are of modulus strictly smaller than 1. This means that the necessary condition is:

$$-1 < 1 - 2\,\mu\,\lambda_p < 1, \quad \forall p,\ 1 \le p \le MK$$

After some manipulation, we obtain the equivalent necesssary and sufficient condition:

$$0 < \mu < \frac{1}{\lambda_{max}}$$

One sees that μ cannot be taken too large. When this condition is satisfied, filter $\mathbf{w}$ converges in mean to the optimal one, $\mathbf{w}_o$. In practice, μ will have to be taken much smaller in order to reach an acceptable variance at the output. The output variance can also be assessed by similar calculations.

3 Exercises

1. **Exponential averaging.** Let $x(n)$ be a discrete-time zero-mean stationary stochastic process, and denote $v = \mathrm{E}\{x(t)^2\}$ its variance. Define the following two estimators of v:

$$\begin{aligned} v_1(N) &= \frac{1}{N}\sum_{i=1}^{N} x(i)^2 \\ v_2(n) &= \alpha\,v_2(n-1) + (1-\alpha)\,x(n)^2; \quad 0 < \alpha < 1,\ v_2(0) = 0. \end{aligned}$$

   Assuming the process $x(n)$ is white, compute (i) the bias and (ii) the variances of estimators $v_1(n)$ and $v_2(n)$. Deduce from this result that estimator v_2 with exponential averaging rate α can be seen, for large n, as an equivalent estimator $v_1(N_\alpha)$ with $N_\alpha = (1-\alpha^2)(1-\alpha)^{-2}$ samples. (iii) What happens when α tends to 1?

2. **Adaptive implementation of the noise canceller.** With the help of the previous exercice, propose another implementation of solution (7).

3. **Capon's spatial filter.** Let $\mathbf{z}(t)$ be the signal received on an array of sensors, and denote $\mathbf{R}$ its covariance matrix. In antenna arry processing, it is often useful to implement what is called the Minimum Variance Distorsionless Response (MVDR) spatial filter, yielding an estimate of the signal coming from direction $\mathbf{a}$, $||\mathbf{a}|| = 1$:

$$s(t) = \frac{\mathbf{a}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{z}(t)}{\mathbf{a}^{\mathrm{T}}\mathbf{R}^{-1}\mathbf{a}}$$

a) Show that the expression of $s(t)$ can be obtained by minimizing the output variance under the (linear) constraint that the spatial filter response is equal to 1 in direction $\mathbf{a}$.
b) Show that it is possible to compute $s(t)$ in real time by using the LMS algorithm. *Hint:* denote $\mathbf{P}$ the projection matrix $\mathbf{P} = [\mathbf{I} - \mathbf{a}\,\mathbf{a}^{\mathrm{T}}]$, and express first the deterministic gradient descent, and then its stochastic version: $\mathbf{w}^{(k+1)} = \mathbf{P}[\mathbf{w}^{(k)} - \mu\,\mathbf{z}(k)\,s(k)] + \mathbf{a}$; $\mathbf{w}^{(0)} = \mathbf{a}$.
c) Show that the deterministic and stochastic iterations are simpler if we assume that $\mathbf{a}^{\mathrm{T}}\mathbf{w}^{(k)} = 1$ for every k. But what may happen if rounding errors accumulate?

References

[1] P. Comon and D. T. Pham, *An error bound for a noise canceller*, IEEE Trans. on ASSP, 37 (1989), pp. 1513–1517.

[2] B. Widrow, J. R. Glover, J. M. McCool, et al., *Adaptive noise cancelling : Principles and applications*, Proc. of the IEEE, 63 (1975), pp. 1695–1716.

[3] V. Zarzoso, R. Phlypo, O. Meste, and P. Comon, *Signal extraction in multisensor biomedical recordings*, in Advances in Biomedical Engineering, P. Verdonck, ed., Elsevier BV, Oxford, UK, 2009, pp. 95–143.