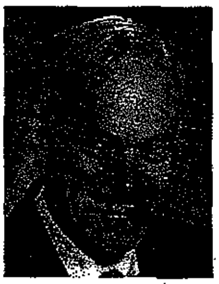
ICIAM 2003

Applied Mathematics Entering the 21st Century

Invited Talks from the ICIAM 2003 Congress

Edited by
James M. Hill
Ross Moore

Contents



Chapter 1

Pulling the Information Out of the Clutter

Brian D. O. Anderson[*†]

Abstract: Fixed and wireless telecommunications systems, sonar systems, navigation devices, image processing algorithms—these are all examples of where signal processing is used. Much signal processing is based on statistical models of processes generating the signals and the contaminating noise. This paper traces the development of statistical processing theories, beginning with Wiener filtering, continuing through Kalman filtering, and ending with Hidden Markov Models. Different assumptions underpin these theories, and also very different mathematics. Yet a number of common features remain.

Contents

[*]Brian D. O. Anderson is Chief Scientist at National ICT Australia, and Professor at Research School of Information Sciences and Engineering, The Australian National University, Canberra.

[†]National ICT Australia is funded by the Australian Government's Department of Communications, Information Technology and the Arts and the Australian Research Council through Backing Australia's Ability and the ICT Centre of Excellence program.

1 Introduction

Much of signal processing is concerned with extracting relevant information from measurements in which that relevant information is contained in some way, but generally buried in noise. Determining the presence of genuine target reflections in a signal picked up by a radar receiver; figuring out what the message is in a fax which is blurred; minimising the effect of extraneous signals for a user of a hearing aid trying to talk to another individual; these, and a whole host of other examples, constitute situations where signal processing is a must.

In the following sections, we will discuss three different approaches to signal processing using statistical ideas, those associated with Wiener filtering [1], Kalman filtering [2], [3], and Hidden Markov Model filtering [4], [5]. A further section is provided on what is termed as smoothing, and this attempts to further illustrate the very real similarities among the different approaches to filtering, despite the huge differences in mathematical tools. The final section summarises key conclusions.

We conclude this section by recording an example drawn from a recent project.

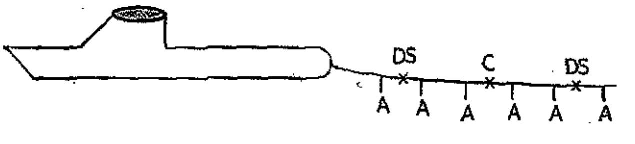


Figure 1. *Submarine with towed array.*

Figure 1 is a diagrammatic representation of a submarine trailing a towed array. A towed array is a cable on which are located a large number of acoustic sensors (labelled with the letter A), and the purpose of the acoustic sensors is to listen for other vessels. Location of acoustic sensors on the towed array rather than location of acoustic sensors on board the submarine allows use of an effectively much bigger and thus more effective acoustic antenna, and lessens difficulties associated with self-noise generated by the submarine. Satisfactory use of the collection of acoustic sensor signals, however, requires knowledge of the shape of the array. The known motion of the submarine together with equations of motion of a towed cable (generally modelled with a nonlinear partial differential equation) allow the generation of an estimate of the shape of the array, but this will be deficient for at least two reasons. First, the equations of motion of the towed cable are only approximations of reality (i.e., there is modelling error); and second, there may well be currents

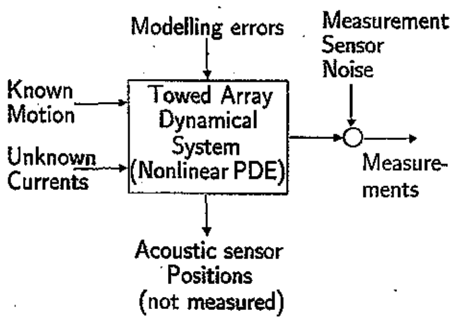


Figure 2. *Abstract representation of array shape estimation problem.*

giving rise to further forces on the array, and thus distortions of it.

For this reason, it is desirable to find techniques for improving the determination of array shape. To this end, one can contemplate including along the array some depth sensors and compasses (labelled DS and C), which provide some sort of (noisy) measurement information relevant to determining the shape of the array.

At this stage, one can represent the situation abstractly via Figure 2. The inputs on the left include the known motion, which is the submarine motion. The rest of the diagram is self-evident.

In order to obtain the acoustic sensor positions, a filter is needed. The signals driving the filter are the measurements from the sensors as well as the motion of the submarine, and, of course, the equations defining the filter in some way will depend on the model of the towed array. Filtering theory in the sense of Wiener [1] and Kalman [2], [3] attempts to provide a technique for determination of a filter, and for predicting the performance of the filter (as measured, for example, by the mean-square error in the estimates produced by the filter).

A description of the towed-array problem can be found in [6].

2 Wiener Filtering

Wiener filtering theory [1] is probably the first attempt to provide optimal filters in situations where signals are characterised by random processes. Indeed, the preface to [1] states: "Largely because of the impetus gained during World War II, communication and control engineering have reached a very high level of development today .... The point of departure may well be the recasting and verifying of the theories of control and communication ... *on a statistical basis.*" Actually, the preface also notes that the work had its origins in ideas of Kolmogorov and Kosulajeff which was published between 1939 and 1941.

A basic situation handled by Wiener filtering is depicted in Figure 3. Figure 3(a) denotes the signal model. The designation 'measurement' is just that: it is what is available to an observer, and can be used for processing. The measurement itself is to be regarded as a sum of signal and noise. The noise, more precisely termed 'measurement noise', is assumed to be a stationary process with

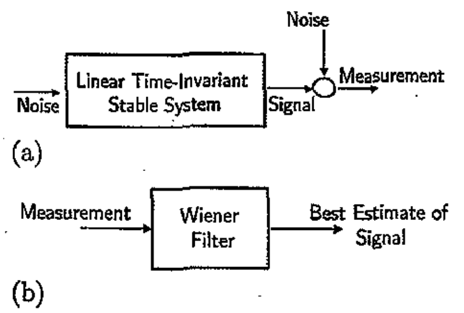


Figure 3. (a) *Signal model;* (b) *basic set-up for Wiener filtering.*

zero mean and known spectrum. The signal is also assumed to be a stationary process. Its generation is modelled through the use of a so-called 'input-noise process', assumed to be zero mean stationary and white[1], passing into a linear time-invariant stable system. Virtually any smooth spectrum can be achieved in this way. The input-noise process and the measurement-noise process are normally assumed to be independent (which implies that the signal and the measurement-noise process are independent). Note that there is no real loss of generality in assuming that the input-noise process is a white process, because if it is not, a 'shaping filter' driven by white noise can be regarded as generating the input process, and then the shaping filter can be combined with the linear, time-invariant stable system between the input-noise process and the signal. The measurement process is often taken as white, but not always.

The Wiener filter, exhibited in Figure 3(b), is the device which reconstructs in real time a 'best estimate' of the signal from the measurement process.

One can either assume that the two noise processes are Gaussian, in which case one can show that the Wiener filter is necessarily linear. Or one can postulate that one is seeking a best estimate among those achievable by linear filters, and not then require that the noise processes are Gaussian. Either way, the Wiener filter is itself linear.

What is meant in this context by the words "constructs in real time a 'best estimate' of the signal from the measurement process"?

Denote the signal by $s(.)$ the measurement by $z(.)$ and the signal estimate by $\hat{s}(.)$ The mean square error at time t associated with the estimate $\hat{s}(t)$ of $s(t)$ is $E[s(t) - \hat{s}(t)]^2$ and it is this quantity which is minimised for the particular choice of filter (a Wiener filter) linking $z(.)$ to $\hat{s}(.)$ The words "in real time" connote that t is a running variable, $\hat{s}(t)$ is available at time t, and necessarily only depends on values of $z(\tau)$ for $\tau < t$ or possibly $\tau \leq t$.

There is of course a theory explaining how the Wiener filter can be calculated from problem data. The central component of the calculation is termed 'spectral factorisation'. With some simplification at the edges, the spectral factorisation

[1] A white process is one with constant spectrum. A gaussian white process is obtainable as the derivative of a Wiener process.

problem looks like the following:

> Suppose $S(\omega)$ is the spectrum of the signal process, and suppose $N(\omega)$ is the spectrum of the measurement-noise process. This means, given independence of the signal and the noise processes, that the measurement process has spectrum $S(\omega)+N(\omega)$. Spectral factorisation involves finding a function $H(s)$ of the complex variable s with the following properties:
>
> (i) $H(s)$ and $H^{-1}(s)$ are analytic in $\mathrm{Re}[s] \geq 0$
>
> (ii) $|H(j\omega)|^2 = S(\omega) + N(\omega)$ for all real ω.

Much of Wiener's work was involved with explaining how to compute the function H. Reference [1] restricts attention, as we have done above, to the case where the signal and noise are scalar processes. In this situation there is in formal terms a formula involving integrals for obtaining H from the spectral information; however, use of approximate integration methods in order to obtain numerical results may be perilous. There is an extension to the vector process case, but not of the formula itself. However, in the event that $S(\omega) + N(\omega)$ is rational in ω, the calculations are hugely simplified; there is even a simple way to deal with the vector process case.

2.1 An Example

By way of illustration, consider Figure 4.

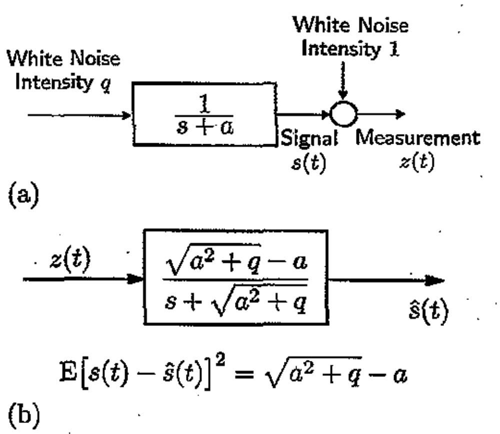


Figure 4. *(a) Signal model; (b) Basic set-up for Wiener filtering.*

To those unfamiliar with the block diagram notation widely used by of electrical engineers, they should understand that Figure 4(a) captures the notion that $\dot{s} + as = w$ where $w(.)$ connotes the input-noise process, here assumed to be white noise with intensity q. This means that the spectrum of w assumes the value q at all frequencies. The spectrum of the process $s(.)$ is $S(\omega) = \dfrac{q}{\omega^2 + a^2}$; then

$$S(\omega) + N(\omega) = \frac{q}{\omega^2 + a^2} + 1 = \frac{\omega^2 + a^2 + q}{\omega^2 + a^2} \tag{2.1}$$

and

$$H(s) = \frac{s + \sqrt{a^2 + q}}{s + a} \tag{2.2}$$

in that

$$\left|H(j\omega)\right|^2 = S(\omega) + N(\omega) \tag{2.3}$$

with H and H^{-1} analytic in $\mathrm{Re}[s] \geq 0$.

Also shown in the figure is the Wiener filter and the value of the mean square error. The block diagram notation is shorthand for

$$\dot{\hat{s}} + \left(\sqrt{a^2 + q}\right)\hat{s} = \left(\sqrt{a^2 + q} - a\right)z\,. \tag{2.4}$$

It turns out that when the measurement noise is white noise of intensity one, the Wiener filter transformation is always $1 - H^{-1}$.

When the spectra are rational, the calculation of H requires in effect simply the determination of the zeros of two polynomials. The numerator and denominator of $S(\omega) + N(\omega)$ are even in ω^2. With the substitution $\omega^2 = -s^2$, the resulting numerator and denominator polynomials in s, call them $m(s)$ and $n(s)$ respectively, are both factored as

$$m(s) = g(s)\,g(-s)\,, \qquad n(s) = h(s)\,h(-s)$$

where all zeros of $g(.)$ and $h(.)$ lie in $\mathrm{Re}[s] < 0$. Then

$$H(s) = \frac{g(s)}{h(s)}\,. \tag{2.5}$$

Obviously, H and H^{-1} are analytic in $\mathrm{Re}[s] \geq 0$, and it is easily checked that $S(\omega) + N(\omega) = \left|H(j\omega)\right|^2$.

3 Kalman Filtering

The Kalman Filter [2], [3] is a variant on the Wiener Filter in several respects. As with the Wiener Filter, we contemplate a signal model and an associated filter. See Figure 5. The key changes between the Wiener Filter and Kalman Filter set-up are as follows:

- The input noise to the signal modal must be white (although this is no real loss of generality, since a device called a 'shaping filter' can be used to cope with a non-white process, at least if it has a stationary, rational spectrum).

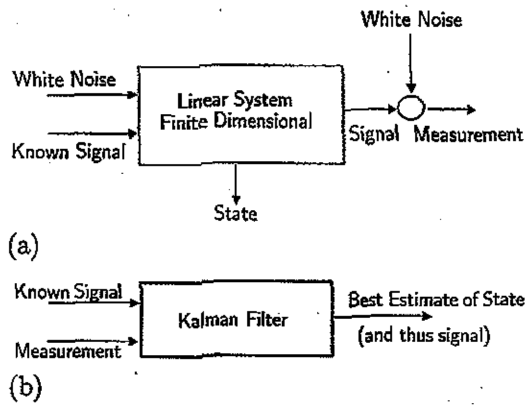


Figure 5. *(a) Signal model; (b) Basic set-up for Kalman filtering.*

- There can be a known, deterministic input in addition to the white-noise process. (The towed array problem is an example where there is a known input signal, the known motion of the submarine.)
- The system linking the input noise to the signal process is necessarily finite-dimensional; i.e., describable by an ordinary differential equation.
- The system linking the input noise to the signal process *is not* assumed to be time-invariant or stable. If however it is unstable, it must be viewed as being switched on at some finite time, rather than the infinitely remote past, to negate the possibility of signals having infinite variance.
- The measurement process is always assumed to be white.
- Quite evidently, the Kalman Filter in order to produce a best estimate of the signal should use both the known input signal and the measurement. But, as it turns out, the Kalman filter seeks not just to estimate 'the signal' but also the whole state of the finite-dimensional linear system sitting between the input process and the signal itself. Obviously, it should be intuitively clear that with a best estimate of the state, one should be able to come up with a best estimate of the signal also.

The great advances in the Kalman Filter are the introduction of possible time-variation in the signal model and non-stationarity in the underlying random processes: the input process and the measurement process need not be stationary; the linear system need not be time-invariant; operations do not have to begin in the infinitely remote past.

The Kalman Filter itself turns out to be a linear finite-dimensional system. It will only be a time-invariant system in the event that all processes in the signal model are stationary.

The calculation of the Kalman Filter is totally different to the calculation of the Wiener Filter. To set it out, we need to formalise some of the above.

3.1 Kalman Filter Equation

The signal model is defined for $t \geq t_0$ by

$$\dot{x} = F(t)\,x + G(t)\,w(t) + \Gamma(t)\,u(t) \tag{3.1}$$

$$s(t) = H(t)\,x \tag{3.2}$$

$$z(t) = H(t)\,x + n(t) \tag{3.3}$$

where $E[w(t)] = 0$, $E[w(t)\,w^{\mathsf{T}}(s)] = Q(t)\,\delta(t-s)$, $E[n(t)] = 0$, $E[n(t)\,n^{\mathsf{T}}(s)] = R(t)\,\delta(t-s)$ with $w(.)$, $n(.)$ independent processes. The initial condition $x(t_0)$ at time t_0 is a random variable independent of $w(.)$ and $n(.)$ with mean $\bar{x}(t_0)$ and variance $\bar{P}(t_0)$. The symbol $u(t)$ denotes the deterministic external input. The matrices $F(.)$, $G(.)$ etc. are of appropriate dimensions, $Q(t) = Q^{\mathsf{T}}(t) \geq 0$ and $R(t) = R^{\mathsf{T}}(t) > 0$ for all t.

The Kalman filter is defined by

$$\dot{\hat{x}}(t) = \left[F(t) - P(t)\,H(t)R^{-1}(t)\,H^{\mathsf{T}}(t)\right]\hat{x}(t) + P(t)\,H(t)R^{-1}(t)\,z(t) + \Gamma(t)\,u(t) \tag{3.4}$$

$$\hat{x}(t_0) = \bar{x}(t_0) \tag{3.5}$$

$$\hat{s}(t) = H(t)\,x(t) \tag{3.6}$$

and $P(t)$ is a symmetric nonnegative definite matrix solving the following matrix Riccati differential equation:

$$\dot{P} = PF^{\mathsf{T}} + FP - PHR^{-1}H^{\mathsf{T}}P + Q\,, \quad P(t_0) = \bar{P}(t_0)\,. \tag{3.7}$$

Among all possible filters where $\hat{x}(t)$ is constructed using $z(s)$ for $s \leq t$, the Kalman filter ensures

$$E\Big\{\left[x(t) - \hat{x}(t)\right]^{\mathsf{T}}\left[x(t) - \hat{x}(t)\right]\Big\} \tag{3.8}$$

is minimised for all t (minimum variance property).

As intimated above, with the right assumptions on the signal model, the Kalman Filter will be time-invariant. Also, in the event that the deterministic signal at the input of the signal model is zero, the signal model for the Kalman Filter and the signal model for the Wiener Filter have the potential to coincide. It follows therefore that there are some circumstances in which the optimal filter can be calculated either with the Wiener approach or with the Kalman approach. Such situations are necessarily those where there is stationarity and finite-dimensionality of the signal model, white constant intensity input noise and white constant intensity measurement noise, and a time of commencement in the infinitely remote past, so that all signals are stationary.

3.2 Example

To illustrate this point, we explain how the Kalman Filter can be calculated for the example of Figure 4. The signal model of Figure 4(a) can be written as

$$\dot{x} = -ax + w\,, \qquad s = x\,, \qquad z = x + n$$

i.e., $F=-a$, $G=1$, $\Gamma=0$, $H=1$, $Q=q$, $R=1$ and initial time is $-\infty$.

The Riccati equation is

$$\dot{P}=-2Pa-P^2+q.$$

It can be verified that for any $\bar{P}(t_0)>0$, when $t_0\to-\infty$, the solution of the Riccati equation is independent of the initial conditions, and is constant, *viz.*,

$$P(t)=\sqrt{a^2+q^2}-a.$$

Accordingly, the Kalman filter is

$$\dot{\hat{x}}=-\sqrt{a^2+q^2}\,\hat{x}+(\sqrt{a^2+q^2}-a)z\,,\qquad \hat{s}=\hat{x}.$$

This is precisely what Figure 4(b) shows.

In situations where either Kalman or Wiener filtering ideas can be used on the same problem, one has stationarity and one has a rational spectrum. It turns out that the Kalman filter approach involves steady-state Riccati equations and there is a deep connection with rational spectral factorisation.

3.3 Time Constants and Exponential Forgetting

There is an important property common to Wiener and Kalman Filters that is actually not universal, but is obtained in most circumstances. Strict theorems are available defining the circumstances under which these properties are obtained; see for example [2], [3], and [7]. There is a notion of a 'time constant' associated with Wiener and Kalman filters. What does this mean?

- Old measurements are forgotten exponentially fast. The best estimate of the state or the signal at a particular time t depends in an exponentially decaying fashion on prior measurements. If measurements sufficiently long ago were inaccurate not just because of the noise but because of sensor failure, this would not affect matters sufficiently far away from the time at which the erroneous measurements were collected.

- Initial state information of the Kalman Filter is forgotten exponentially fast. To understand this statement, recall that at some time (which might be in the infinitely remote past) the Kalman Filter has to be turned on. There is a best initial state for the Kalman Filter, namely, the mean assumed for the initial condition of the signal model. Obviously, one must reckon with the possibility that a highly inappropriate initial state for the Kalman Filter is selected. The point of the observation is that any damage caused by the inappropriate selection will be forgotten exponentially fast.

- Round-off and similar errors can only accumulate to a limited extent. Suppose that the Wiener or the Kalman Filter is implemented on a computer, so that at every step in the calculation some little error is introduced. One should have an *a priori* concern as to whether these errors will accumulate in an ultimately damaging way. The point of the remark is that this will not happen.

The array shape estimation problem introduced in Section 1, see Figure 2, was tackled in [6] using Kalman Filtering ideas. A key step in the modelling is to replace the non-linear partial differential equation model of the towed array by a linear-system model. Also, the replacement model must be finite-dimensional. Thus, as with many practical problems, the initial mathematical problem has to be approximated or modified, to suit the Kalman Filtering framework. The acoustic sensor positions correspond to some components of the state vector in the signal model and their estimates to some components of the state vector in the Kalman Filter. The approximation error in replacing the Figure 2 model is assumed to be swept up in some way by the incorporation of noise signals in the signal model; i.e., the noise signals in the signal model are meant to capture not just genuine noise from sensors, or the uncertainty associated with currents, but also (obviously in a very crude fashion) the inaccuracies associated with the approximation inherent in the modelling process. That these inaccuracies will not overwhelm the calculations as they evolve in time is also a consequence of the time-constant/exponential-forgetting concept described above.

4 Hidden Markov Models

Wiener and Kalman filtering theories are concerned with filtering of signals and linear systems. The theory can be pushed to consider some levels of nonlinearity, typically when linearisation is applicable; but in no sense do the theories provide a general theory for the filtering of nonlinear systems. There is however a theory which can capture many nonlinear filtering problems, and that is based on Hidden Markov Models [4], [5].

As noted in the abstract of [5], Hidden Markov Models (HMMs) were initially introduced in the late 1960s and early 1970s (i.e., about 30 years ago), and their popularity has slowly grown. To quote from the abstract of [5]:

> "There are two strong reasons why this has occurred. First, the models are very rich in mathematical structure and hence can form the theoretical basis for use in a wider range of applications. Second the models, when applied properly, work very well in practice with some important applications."

The reasons might just as well have been advanced for the Wiener filter and Kalman filter. But for HMMs, the mathematical structure is very different, and the successful applications are also very different. The mathematical structure does not include spectral factorisation or Riccati equations. But it does include the theory of positive matrices (to be distinguished from positive-definite matrices), including (as work of this decade has revealed) a form of time-varying extension of Perron–Frobenius theory, [8].

4.1 Examples of Hidden Markov Models

Before defining what a Hidden Markov Model is, let us give several examples. The first is a very old one, the random telegraph wave (see Figure 6).

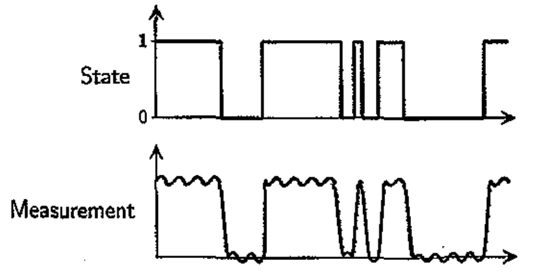


Figure 6. *Noisy measurement of a random telegraph wave.*

One assumes that a signal is transmitted which takes the value zero or one. One has available noisy measurements of that signal, and from the noisy measurements, one is required to reconstruct the original signal. (The common noise model is additive white gaussian noise.) The transitions within the original signal are assumed to occur in a Poisson manner, and the number of levels in the original signal (here two) can be generalised to be finite, but not to be infinite.

For the second example, consider the problem of listening from one submarine for the engine of another submarine. One can postulate that the engine speed of the other submarine has a fundamental frequency lying in one of a finite set of frequency ranges, and the transition probability for movement from one range to another is known. Noisy estimates (in effect noisy measurements) are available of the particular range in which the fundamental frequency of the other submarine's engine lies, and the problem is to properly reconstruct the activity of the other submarine's engine from the estimates.

For a third example, consider a significantly nonlinear variant on the signal model of Figure 4(a). In particular, suppose that replacing the equation

$$\dot{x} = -ax + w \tag{4.1}$$

there appears the equation

$$\dot{x} = -a(x) + w$$

where $a(.)$ is some nonlinear function of x. One cannot expect Kalman or Wiener filtering to work well, unless perhaps, and then only perhaps, $a(.)$ is close to being linear. To cope with the nonlinear case, we could imagine partitioning up the real axis on which x lies into a finite set of intervals (two of which would be semi-infinite), and consider transitions in the variable x from one region to another. When x is in the i^{th} region, we could regard that as equivalent to the i^{th} state of a Hidden Markov Model with a finite number of states being the current state of the HMM. The finer the sub-division, the more accurately would the HMM capture the model with continuous x. This is the way in which Hidden Markov Models can be used to cope with problems which are too nonlinear for Kalman filters, where it happens to be true that some linearisation can be acceptable.

4.2 Formal Description with Finite-State Hidden Markov Model

In this sub-section we shall try to capture in a more abstract framework the contents of the previous examples. We shall assume that the underlying time process is a discrete time one rather than continuous time, which eases the exposition very substantially. The use of discrete-time modelling in signal processing has become increasingly common, driven not just by questions of ease of exposition, but the very digital nature of much of the hardware which serves as the implementation platform.

There is an underlying state process $\{X_0, X_1, \ldots, X_k, \ldots\}$ and X_k (the process at time k) can assume one of a finite set of values, for convenience $1, 2, \ldots, N$. The quantity $\Pr[X_{k+1} = i \mid X_k = j] = a_{ij}$ is a transition probability, and X_k is a Markov process. We denote by A the matrix $[a_{ij}]$. There is also an output process $Y_0, Y_1, \ldots$. We shall assume that Y_k takes one of a finite set of values, for convenience labelled $1, 2, \ldots, M$. (There are many examples where Y_k assumes continuous values; for example, when it is equal to gaussian noise plus the state. However, for the sake of this paper we shall assume the simpler case of a finite set of values. This finite set incidentally could arise through quantisation of a continuum.) The link between the state process and the output process is defined by $\Pr[Y_k = m \mid X_k = n] = c_{mn}$ and we denote by C the matrix (c_{mn}).

Evidently, then, two matrices whose entries are all probabilities, A and C, describe the Hidden Markov Model process.

4.3 Hidden Markov Model Filter

An HMM filter is, in quite precise terms, a device for calculating the N-vector whose i^{th} entry is $\Pr[X_k = i \mid Y_0, Y_1, \ldots, Y_k]$. This means that the filter uses the measurements up to time k to provide the best possible statement concerning the knowledge of the state at time k. For the simple HMM setup that we have described, it is fairly easy to obtain filtering equations by straightforward application of Bayes' Theorem. The update process involves two steps, incorporating a time update of the state variable with no extra measurements, and then adding in the extra measurement associated with an update. More precisely, let $\Pi_{k|k}$ be the vector with i^{th} entry $\Pr[X_k = i \mid Y_0, \ldots, Y_k]$, and $\Pi_{k+1|k}$ the vector with i^{th} entry $\Pr[X_{k+1} = i \mid Y_0, \ldots, Y_k]$. Then

$$\Pi_{k+1|k} = A\,\Pi_{k|k} \tag{4.2}$$

$$\Pi_{k+1|k+1} = \frac{1}{[1 \ldots 1]\, C_{Y_{k+1}} \Pi_{k+1|k}}\, C_{Y_{k-1}} \Pi_{k+1|k}\,. \tag{4.3}$$

4.4 The Forgetting Property

At this stage one can ask similar questions to those which can be asked regarding Wiener and Kalman filters:

Are old measurements forgotten, is an inappropriate filter initialisation forgotten, and are round-off and similar errors guaranteed not to overpower the calculation?

As for Kalman and Wiener filtering problems, the answer is, in general, *yes.* The qualification is one which can be expressed in technical terms [9]–[13]; in broad terms it demands that filtering problems be well-posed. The general conclusion is in fact that *there is an exponential forgetting property,* just like that for Kalman and Wiener filtering. Incidentally, obtaining these conclusions for continuous-time Hidden Markov Models is much more difficult technically.

There is a new angle here, which does not arise in Kalman and Wiener filters, and it should be noted. What we have just said is that the calculations leading to the conditional probability associated with the filtering problem are ones in which an exponential forgetting property is found. Suppose that one focuses on the actual production of a state estimate. Thus, one could logically define a filtered state estimate by saying that $\hat{x}_{k|k} = J$ if J maximises $\Pr\left[X_k = i \mid Y_0, \ldots, Y_k\right]$ over i. Then it turns out that $\hat{x}_{k|k}$ is determined with a *finite memory*; i.e., $\hat{x}_{k|k}$ depends on $Y_k, Y_{k-1}, \ldots, Y_{k-l}$ for some fixed l and all k. At this stage, theory is not available to estimate l easily [14].

4.5 Mathematics of the Exponential Forgetting Property

The exponential forgetting property can be established by an elegant extension of the Perron–Frobenius theory on the eigen-properties of matrices of non-negative or positive entries, [8].

Consider equations 4.2 and 4.3, and the following two equations:

$$\Sigma_{k+1|k} = A\,\Sigma_{k|k} \tag{4.4}$$

$$\Sigma_{k+1|k+1} = C_{Y_{k+1}}\,\Sigma_{k+1|k}\,. \tag{4.5}$$

The effect of the scalar division on the right side in 4.3 and absent in 4.5 is to normalize; i.e., scale the right side of 4.3 so that the vector entries add up to 1. This means that 4.4 and 4.5 together constitute an unnormalized version of 4.2 and 4.3. Thus the behaviour of 4.4 and 4.5 will be able to predict, in many ways, the behaviour of 4.2 and 4.3.

Now observe that 4.4 and 4.5 can be combined together to give

$$\Sigma_{k+1|k+1} = C_{Y_{k+1}} A\,\Sigma_{k|k}\,. \tag{4.6}$$

The matrix $C_{Y_{k+1}}$ can only assume one of M values and so we could rewrite 4.6 as

$$\Sigma_{k+1|k+1} = D_k\,\Sigma_{k|k}\,, \tag{4.7}$$

where D_k is drawn from a finite set of known matrices, call them $A_1, A_2, \ldots, A_M$. The extension of the Perron–Frobenius theorem is found in [8]. Let $\{A_1, \ldots, A_M\}$ denote a finite set of matrices with positive entries, and let $E_N = D_N D_{N-1} \cdots D_1$, where $D_i \in \left[A_1, \ldots, A_M\right]$. Then as $N \to \infty$, $E_N \to \mu_N \nu^{\tau}$ for variable vector μ_N and some fixed vector ν, exponentially fast. (The Perron–Frobenius theorem deals with the case where all the D_i are the same.) There are incidentally extensions of this inhomogeneous product result to cope with non-negative matrices, and such extensions can be useful for the application to HMMs.

The above result implies

$$\Sigma_{N|N} \longrightarrow \mu_N \, \nu^{\mathrm{T}} \, \Sigma_{0|0}$$

exponentially fast. Observe that different values of $\Sigma_{0|0}$ will lead to different values of $\nu^{\mathrm{T}} \Sigma_{0|0}$, which is a scalar; i.e., $\Sigma_{0|0}$ only affects the scaling of $\Sigma_{N|N}$.

If there is normalization, as there is in calculating $\Pi_{N|N}$ from $\Sigma_{N|N}$, $\nu^{\mathrm{T}} \Sigma_{0|0}$ drops out completely. Equivalently, the initial condition is forgotten exponentially fast. Similarly, one argues that old measurements are forgotten exponentially fast.

4.6 Rapprochement between HMMs and Wiener–Kalman Theory

One can pose the question:

> Are there situations in which the Hidden Markov Model approach and the Kalman filter approach or the Wiener approach overlap?

If by the HMM approach, one is talking about finite-state processes, the answer is *no*. However, one can usually regard the type of signal model which appears in a Kalman filter or Wiener filter problem as a limiting version of a Hidden Markov Model signal model. Unpublished work of R. L. Streit has demonstrated that as the limit of the number of states in a certain finite-state Hidden Markov Model is allowed to become infinite, the HMM converges in a certain sense to a Kalman filter signal model, and one can establish convergence of the associated filters as well.

5 Smoothing

In this section, we aim to explain a variant on filtering which applies to each of the three types of filtering we have described.

Any processing of the measurements is often described by the generic term 'filtering'. However, one can particularize the meaning of the word filtering, to distinguish it from a related concept called 'smoothing'. This is the point of this section. Also, we will record some distinctions in the properties of filters and smoothers.

5.1 The Difference between Filtering and Smoothing

Consider Figure 7. For convenience, we shall explain the smoothing concept in terms of a Kalman filtering problem. The explanations carry over to Wiener and HMM problems, and to discrete-time formulations.

This figure depicts one entry of the unknown true state of the system on which the filtering is being performed: it depicts the measurements taken at the output of that system and it depicts one entry of the filtered estimate of the state, obtained at the output of the Kalman filter. The notation $\hat{x}(t \mid t)$ is used to denote a filtered

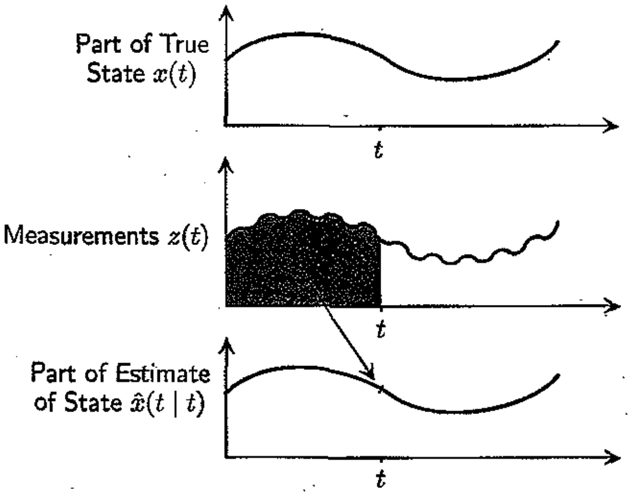


Figure 7. *Representation of filtered-estimate dependence on measurements.*

estimate. The first occurrence of t signifies that we are achieving this estimate of the state at time t and the second occurrence of t signifies we are using measurements occurring up to a time t.

It is intuitively obvious that measurements received after time t must contain some sort of information about $x(t)$. If those measurements are in some way usable, we ought to be able to obtain an improved estimate of $x(t)$; i.e., one with lesser mean-square error. Figure 8 illustrates the distinction between filtering and 'smoothing', where in smoothing we are using measurements not just up to time t but up to some later time T, in order to produce our estimate of $x(t)$. The new estimate is termed $\hat{x}(t \mid T)$.

Leaving aside for the moment the question of how exactly such an estimate might be constructed, it is important to realize that there is one key disadvantage of working with a smoothed estimate; namely, the estimate is not available in real time but only with some delay. For a control application, this may be a fatal disadvantage. However, if one is analyzing what happened in an experiment subsequent to that experiment, there may be no disadvantage at all.

For the sake of completeness, we should mention also the concept of prediction using measurements $z(t)$ up until time t. One seeks to estimate not $x(t)$ but $x(t+\delta)$ for some $\delta > 0$. (The quantity δ may be fixed and t a running variable.) Such an estimation may be relevant in, for example, a rendezvous problem with a moving target with which a rendezvous is sought at a future time. If one can do filtering, it is generally very easy to do prediction; we will devote almost no attention to it.

There are in fact several different types of smoothing which need to be distinguished. These are termed fixed-interval smoothing, fixed-point smoothing, and fixed-lag smoothing.

In fixed-lag smoothing, t is variable and T is set to equal $t + \Delta$, with a fixed

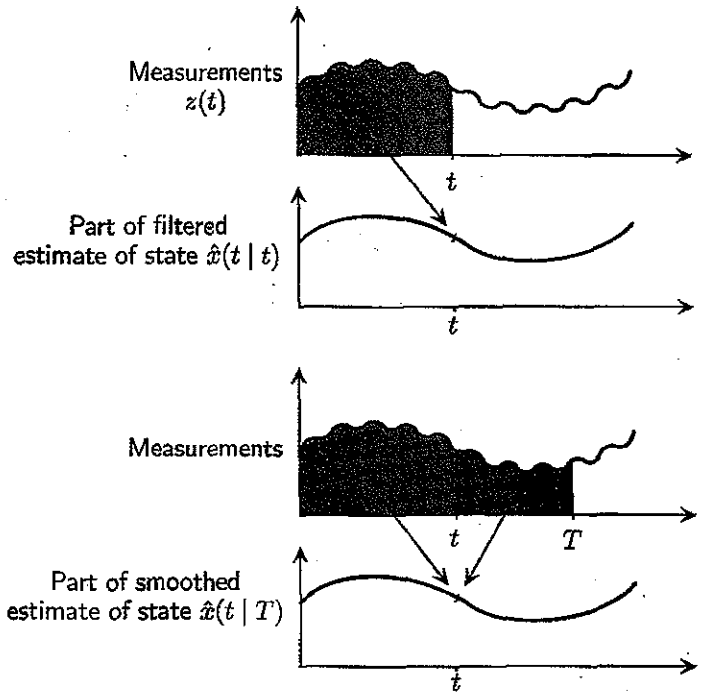


Figure 8. *Contrast of smoothed and filtered estimate dependence on measurements.*

quantity Δ termed the 'lag'. Thus fixed-lag smoothing is like filtering with delay. Figure 9 illustrates how various measurements give rise to a fixed-lag estimate at different time instants.

Fixed-lag smoothing is treated in the Wiener filtering context in [1], in the Kalman filter context in [7], and in the HMM context in [13], with a precursor in [15]. Figure 10 depicts traces of the state of a discrete-time system, a filtered estimate of that state and a fixed-lag estimate if that state together with error performance of the filter and the smoother. An inspection by eye suggests the greater accuracy of the fixed-lag estimate, and this is confirmed by a calculation of the error variance.

In relation to the towed array problem, it is evident that in filtering, measurement information up to a time t would allow estimates of the acoustic-sensor positions at time t and allow listening for other vessels using those sensor estimates. Smoothing would allow a better estimate of acoustic-sensor positions, and allow better listening — but there would be a delay.

5.2 A Mathematical Road Block

Aside from the (admittedly modest) movement in conceptual complexity in passing from filtering to smoothing, there were practical problems associated with the implementation of smoothers. These can best be understood by considering the same

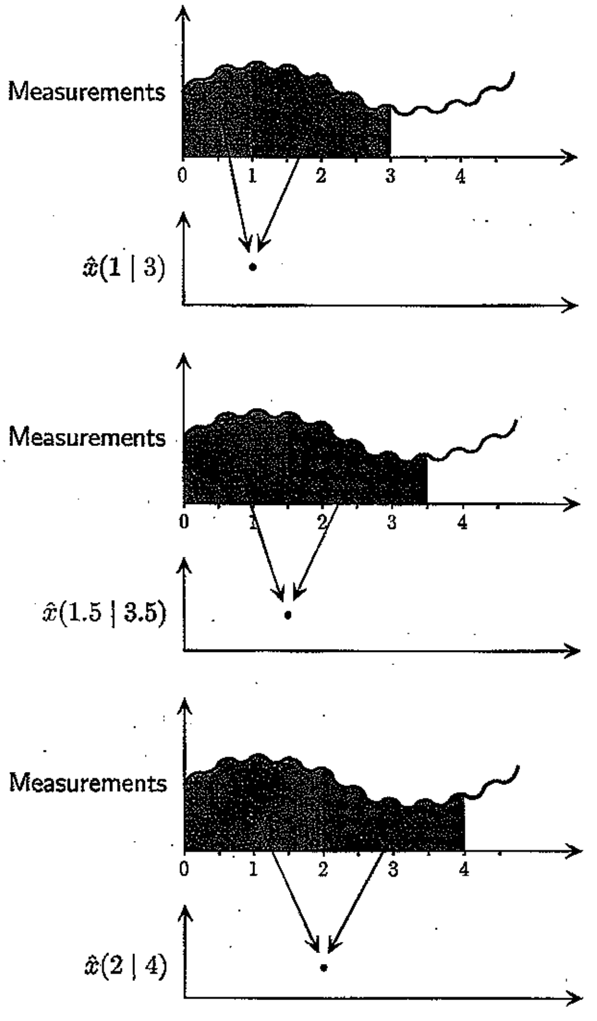


Figure 9. *Representation of generation of a fixed-lag estimate.*

simple example as earlier, illustrated in Figure 4.

With a fixed lag of Δ, the fixed lag estimate $\hat{s}(t-\Delta \mid t)$ can be shown to be given by

$$\hat{s}(t-\Delta \mid t) = \hat{s}(t-\Delta \mid t-\Delta) + (b-a)\frac{\exp(-s\Delta)-\exp(-b\Delta)}{s-b}\left[z(t)-\hat{s}(t \mid t)\right] \quad (5.1)$$

where $b=\sqrt{a^2+q}$ and we have mixed Laplace transform notation with pure time-domain quantities. It turns out that the hardware implementation of a device

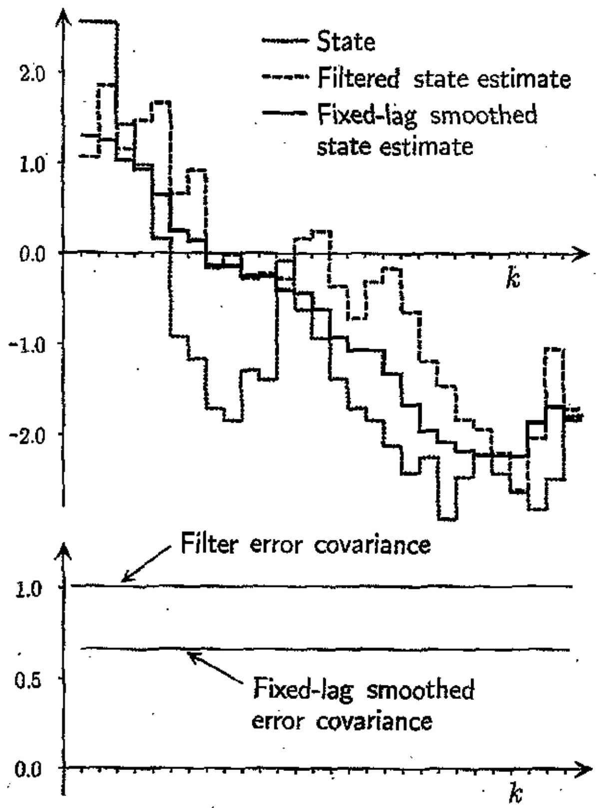


Figure 10. *Simulation data for filtering and fixed-lag smoothing comparison.*

corresponding to

$$\frac{\exp(-s\Delta) - \exp(-b\Delta)}{s-b} \tag{5.2}$$

is apparently straightforward, but this is not actually the case.

More precisely, the "obvious" implementation to an electrical engineer is fatally flawed, due to the inclusion of a guaranteed instability. The "obvious" hardware device has to capture the equivalent of

$$\dot{y} - by = u(t-\Delta) - \exp(-b\Delta)\, u(t) \tag{5.3}$$

[with $u(\,.\,)$ an input and y an output]. An exact solution of this equation results in a bounded mapping from $u \in L_p$ to $y \in L_p$ for any $p \in [1, \infty)$ (with an appropriate initial condition). Nevertheless, the associated homogenous equation

$$\dot{y} - by = 0 \tag{5.4}$$

has exponentially divergent solutions, and any hardware (including software algorithm) based on direct solution of the forced equation will be overpowered by the instability. Fortunately, the problem does not occur in discrete time. (To explain this would lead us too far afield.)

For information on these points, see [16],[17],[18] for two approaches to circumventing the instability above (which is generic, and not particular to this example), [7] for discrete time Kalman filters, and [13] for discrete time HMM filtering.

5.3 Comparative Advantages of Smoothing over Filtering

We have already referred to the key disadvantage of using smoothing as opposed to filtering, including fixed-lag smoothing. This is the delay in obtaining an estimate. The key advantage is the greater accuracy in the estimate. This naturally raises the question: "What improvement can we expect?" A subsidiary question is: "How much lag should one use in fixed-lag smoothing to capture all the significant improvement?"

These questions have been addressed for Wiener filtering problems in a number of papers: see e.g., [19]–[22]. The first key conclusion is that at *high signal-to-noise ratios smoothing gives greater improvement over filtering than at low signal-to-noise ratios.* Denote by P_s and P_f the mean square error in estimating the signal with a smoother and with a filter, respectively. Then [19] provides for a significant family of systems, a bound for the minimum possible value of P_s/P_f in terms of the maximum signal-to-noise ratio. The bound is depicted in Figure 11.

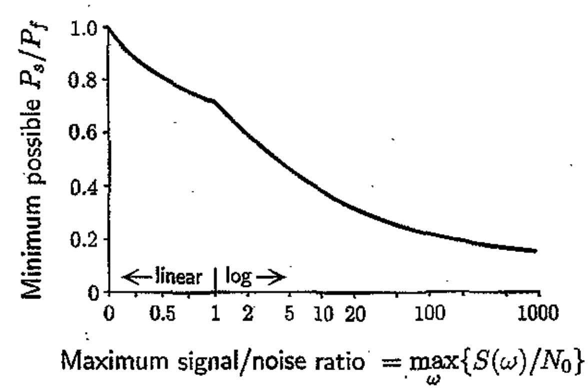


Figure 11. *Smoothing improvement against maximum signal-to-noise ratio.*

At low signal ratios then it is impossible to get much improvement. Note also that the curve does not guarantee that at high signal-to-noise ratios, there has to be a lot of improvement. It simply indicates that there may be a lot of improvement. Nevertheless, examples supported in various references testify to the conclusion that at high signal-to-noise ratios, a significant improvement can be expected.

(Notice that the SNR can go to infinity either as signal power goes to infinity or noise power goes to zero. In the latter case P_s and P_f, both goes to zero and in the limit, the issue of improvement is irrelevant. However for high signal-to-noise ratios, improvement may nevertheless be very desirable: modern digital communications systems after all do seek extremely low error rates.)

The same qualitative conclusions hold for HMM filters. Experimental data was obtained in [15]. The theoretical underpinnings however are not yet complete, as there are no nice formulae for the error measure of an HMM filter. Nevertheless, for the high signal-to-noise ratio case, analytic justification was obtained in [23].

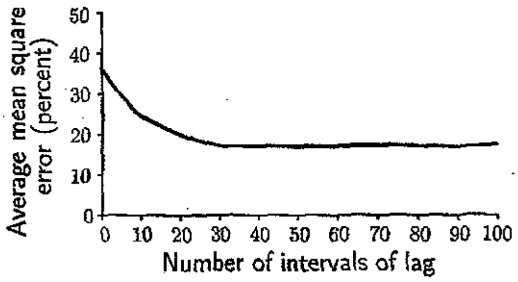


Figure 12. *Variation of smoothing performance with lag.*

A second key conclusion relates to the choice of Δ for fixed-lag smoothing. If Δ is taken to be several times the dominant time-constant of the Wiener or Kalman filter, then one will obtain all the practical improvement that it is possible to obtain using fixed-lag smoothing. A typical curve illustrating the situation is shown in Figure 12.

The case of zero lag corresponds to filtering. As the lag is increased, the mean square error goes down monotonically in fact but the benefit from further increases in lag gradually tails off (in fact it tails off exponentially) until a lag is reached at which further increase of Δ is pointless.

For HMMs, experimental evidence appeared in [15], but it took approximately 25 years before the results could be explained, in [13].

6. Conclusion

In this paper, we have tried to sketch some of the parallels between Wiener, Kalman and HMM filtering, parallels which persist despite very different styles of mathematics. The key results we have chosen to highlight concern:

(1) exponential forgetting properties in filtering, with the relations of the associated time-constant to the understanding of lag adjustment in fixed-lag smoothing;

(2) the benefit of fixed-lag smoothing over filtering being much greater for low-noise situations.

While the details of (1) are well worked out, this is not the case for (2). The absence of error-rate formulae for HMMs is one roadblock.

One would imagine the ideas would be relevant in other domains; e.g., Image Processing.

Bibliography

[1] N Wiener, *Extrapolation, interpolation and smoothing stationary time series*, MIT Press 1949.

[2] R. E. Kalman, R. S. Bucy, *New results in linear filtering and prediction theory*, J. Basic. Eng. Trans. ASME Series D Vol 83 (1961), pp. 95–108.

[3] R. E. Kalman, *A new approach to linear filtering and prediction problems*, J. Basic. Eng. Trans. ASME Series D Vol (1960), pp. 35–45.

[4] R. J. Elliott, L. Aggoun, J. B. Moore, *Hidden Markov Models: estimation and control*, Springer-Verlag 1994.

[5] L. R. Rabiner, *A tutorial on Hidden Markov Models and selected applications in speech recognition.*, Proc. IEEE Vol 77 (1989), pp. 257–285.

[6] A. D. Grey, B. D. O. Anderson, R. R. Bitmead, *Towered array shape estimation using Kalman filters Part I — Theoretical Model*, J. Oceanic Engineering Vol 18 (1993) pp. 543–556.

[7] B. D. O. Anderson, J. B. Moore, *Optimal Filtering*, Prentice–Hall Inc., 1979.

[8] E. Seneta, *Non-negative matrices and markov chains*, 2nd ed. Springer-Verlag, 1981.

[9] B. D. O. Anderson, *New development in the theory of positive systems*, in Systems and Control in the Twenty-First Centrury, C. I. Byrnes, B. N. Datta, C. F. Martin, D. S. Gilliam, eds., Birkhauser, Boston, 1997.

[10] F. LeGland, L. Mevel, *Geometric ergodiety and selected applications in speech recognition*, Proc. IEEE, Vol 77 (1989), pp. 257–285.

[11] A. Arapostathis, S. I. Marcus, *Analysis of an identification algorithm arising in adaptive estimation of Markov chains*, Math. of Control, Signals and Systems, Vol 3 (1990), pp. 1–29

[12] L. E. Baum, T. Petrie, *Statistical inference for probabilistic functions of finaite state Markov chains*, Annals of Math. Stats Vol 37 (1966), pp. 1154–1567.

[13] L. Shue, B. D. O. Anderson, S. Dey, *Exponential stability of filters and smoothers for hidden Markov models*, IEEE Trans. Signal Processing, Vol 46, No. 8, August 1998, pp. 2180–2194.

[14] B. D. O. Anderson, *Forgetting properties for hidden markov models*, in Defenece Applications of Signal Processing, Proc. of the US/Australia Joint Workshop on Defence Applications of Signal Processing, D. Cochran, W. Moran, L. B. White (eds), Elsevier, Amsterdam, The Netherlands, 2001, pp. 26–39.

[15] D. Clements, B. D. O. Anderson, *A nonlinear fixed-lag smoother for finite-state Markov processes*, IEEE Trans. Inform. Theory, Vol IT–21 (1975), pp. 446–452.

[16] S. Chirarattanon, B. D. O. Anderson, *Outline design for stable, continuous-time processes*, Electronic Letters, Vol 8 (1972) pp. 163–264

[17] S. Chirarattanon, B. D. O. Anderson, *Stable fixed-lag smoothing of continuous-time processes*, IEEE Trans. Infor. Theory, Vol IT–19 (1973), pp. 25–36.

[18] P. K. S. Tam, J. B. Moore, *Stable realization of fixed-lag smoothing equations for continuous-time signals*, IEEE Trans. Auto. Control, Vol AC–19 (1974), pp. 84–87.

[19] S. Chirarattanon, B. D. O. Anderson, *Smoothing as an improvement on filtering: a universal bound*, Electronic Letters, Vol 7 (1971), p. 524.

[20] B. D. O. Anderson, *Properties of optimal linear smoothing*, IEEE Trans. Auto. Control, Vol AC–14 (1969), pp. 114–115.

[21] B. D. O. Anderson, S. Chirarattanon, *New linear smoothing formulas*, IEEE Trans. Auto. Control, Vol AC–17 (1972), pp. 160–161.

[22] J. B. Moore, K. L. Teo, *Smoothing as an improvement on filtering in high noise*, Systems and Control Letters, Vol 8 (1986), pp. 51–54.

[23] L. Shue, B. D. O. Anderson, F. De Bruyne, *Asypototic smoothing errors for hidden Markov models*, IEEE Trans. Signal Processing, Vol 28, No. 12, Dec 2000, pp. 3289–3302.