January, 1982

LIDS-P-1168
(revised)

Some Solutions, Some Problems, and Some Questions

by

Alan S. Willsky

Department of Electrical Engineering and Computer Science
and
Laboratory for Information and Decision Systems
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139

ABSTRACT

This is a slightly expanded written version of a plenary lecture given at the 20th IEEE Conference on Decision and Control, December 1981, San Diego, Calif. This talk is intended to help spark discussion among researchers in decision, control and estimation about the status of research in the field and of promising new directions for research. One such direction, model-based signal processing, is treated in some depth together with an example, an object detection and estimation approach to tomographic reconstruction. Several additional topics are treated through the posing of questions which hopefully will provoke responses.

This work was supported in part by the National Science Foundation under Grant ECS-8012668

Some Solutions, Some Problems, and Some Questions

Alan S. Willsky

Department of Electrical Engineering and Computer Science
and
Laboratory for Information and Decision Systems
Massachusetts Institute of Technology

Introduction

It was a great pleasure and honor for me to have been asked to be a plenary speaker at the Conference on Decision and Control. When Abe Haddad and Bill Perkins asked me to speak to the attendees of the conference, I accepted gladly but with some hesitation, as I felt that I had something important to say, but I knew that it would be very difficult to say it. Shortly after accepting the invitation I made things even more difficult for myself by selecting a "cute" title. While it is possible to be "cute" in 25 words or less, it is much more difficult and most certainly undersirable to be "cute" for an entire hour. Hopefully you will not find this lecture to be too "cute." Hopefully you will not find it to be too homely either.

Now, about the title. To understand what I meant in my title, I think it is worthwhile to visualize it in much the same way I have.

Consider Figure 1[†] which presents an image of the title which is perhaps the first one that springs to mind--three separate phrases, listed with no obvious emphases, although the ordering I've chosen may seem a bit strange. Each phrase conjures up a mental image but it is a relatively amorphous one which could be shaped in a variety of ways.

Without attempting to define these terms directly, let me try to give them a bit more shape through several other images. Intuitively one can think of the research process as consisting of the continuing, dynamic interaction of the three elements listed in my title (Figure 2). Mathematical problems are formulated motivated by specific classes of applications and by the methods of solution that one already knows. Solutions are produced. Throughout this process questions are asked: How can we extend existing mathematical methodologies? How can we use existing methodologies in the context of a specific physical problem to obtain a tractable formulation which addresses the issues of interest in the more ill-defined physical problem? Why are we working on the problem we're working on? Will they increase the scope of theory or theoretical understanding? Or do they meet the specific needs of a class of physical problems? What types of problems should we be attempting to formulate, motivated either by theory or by practice?

It is this interaction of the solving, formulating, and questioning and critiquing of problems that is at the heart of the creative aspects of research. Indeed, in a discipline such as ours, which is not married to a specific application area (which would provide limits and a clear focus for

[†] Figures 1-4 and 9-13 were drawn by my wife Susanna Natti, who, when she is not illustrating esoteric talks such as this, illustrates children's books.

questions) or to a specific set of techniques (which would do the same for the methods of solution and types of problems considered) this interaction is absolutely vital. It is, however, an unfortunate but completely understandable fact that what gets discussed at meetings such as this are, for the most part, solutions. Consequently the image one is more likely to see at a conference is something like Figure 3. The reasons for this are relatively clear. First of all, we all want to let everybody know what we've solved. Secondly, in any relatively homogeneous meeting in which many of the perspectives for motivating, formulating, and solving problems are shared by many of those in attendance, there is less reason to spend significant amounts of time on discussions and critiques of methods and formulations and consequently a danger that no time is so spent. Third, it's difficult to discuss problems and questions as they're far more amorphous. I speak from recent experience on that point thanks to my labors in preparing this talk. Fourth, to be fruitful, such discussions must be completely open and honest critical evaluations of present research directions, on why and how they have evolved in the fashion they have, and where they ought to be going. Given the natural reticence that we each have toward wandering too far from what we know well when engaged in discussions at meetings such as this and the feeling of having to defend or promote our personal research, it is very difficult to have truly constructive discussions.

Just as the reaons are clear for the difficulty in having discussions of problems and questions, the reaons (illustrated in Figure 4) are also clear for the importance of such discussions. In the first place there is the distinct danger of leaving questions out in the cold with solutions, that is the mdethods for solving problems that we have used with great success in the past, providing the primary driving force inthe formulation of problems (note that "Solutions" is leading in the figure). Research that is driven by such a mechanism runs the risk of heading off in a tangential and sterile direction. Furthermore, there is the potential danger that the original motivation for a particular research direction may be lost several generations downstream with younger researchers caught up in research whose purpose they have never really questioned. [Note that professional generations are quite short (I've been a "grandfather" for a number of years now), and thus it doesn't take long for tangents to form]. Finally, by not asking questions, we may miss attractive and exciting areas in which we as a group can make significant contributions.

The key phrase in the preceding sentence is "we as a group." Individuals and small groups working on the same topic certainly do consider questions internally. However it is useful and important to share this process of questioning with the community as a whole. Furthermore, by no means do I wish to imply that individuals question the "relevance" of their work (in my opinion there are real dangers with that as well). Rather, in a field of our size the aggregate energy of the efforts of so many energetic people creates a great deal of momentum, and it's a good idea every once in a while to take a peek out the window of the locomotive to see where it's heading.

Obviously the one-way nature of a lecture like this is inappropriate for such discussion. Also, all it is that I can offer is the perspective and biases of one individual, while the process I've described must involve the full range of views and experience to be found among researchers in our field. What I hope to do, however, is to get some reaction. To accomplish this, what I will do first is to describe in more depth one research direction in which I am personally involved and which I feel is extremely exciting and promising. Following this I'll pose a few questions that I hope will provide further stimulation for introspection, real dialogue, and spirited debate among professionals in the field.

Model-Based Signal Processing

In my opinion one of the most exciting and challenging areas of research is signal processing. Clearly this is not a new topic. Signal processing is an enormous field which has drawn on the expertise possessed by researchers in many disciplines, including ours, for years. However, I feel that we can play an even larger role. To help focus the discussion of why and how we can have an impact in signal processing, I will first make a few general observations that have helped shape my own personal slant toward research in signal processing. Then, a digression -- an example of a particular signal processing problem to illustrate my points. Finally, I'll step back from the example to extract several essential points.

First, a few observations:

1. Signal processing problems of enormous complexity abound. Of course they always have, but now people are seriously entertaining implementing extremely complicated algorithms. For example, three-dimensional seismic imaging is an important topic being explored by oil companies at the present time. Oceanographic signal processing is another example, as is the processing of remotely sensed data. In addition, there are numerous military signal processing problems ranging from correlation guidance and terminal homing, to the processing of sonar data for the detection and tracking of submarines, to the use of ground acoustic data to detect and track objects and vehicles. In all of these applications the ultimate goal is to develop methods for extracting, in a useful form, every last drop of information.
2. In many of these complex signal processing problems the number of degrees of freedom--i.e., the information to be extracted--has a hierarchical structure. For example in three-

dimensional seismics, at the lowest level one may want a complete picture of the subsurface. If one discretizes in each of three orthogonal directions the resulting number of unknowns is enormous. At a higher level perhaps what one wants is to know where the major boundaries are between subsurface layers and where any faults might be located. The corresponding number of degrees of freedom is far smaller.

3. In many of these problems one has a significant amount of a priori information about the phenomenon generating the signals. For example, a great deal is known about ocean temperature variations which could be useful in processing data in order to map sound velocity variations in a region of the ocean.

4. A not insignificant portion of the processing presently done in complex problems makes inadequate (or, in some cases, I would argue, no) use of the available a priori information. In addition, in some cases the hierarchical nature of the information to be extracted is ignored. By this I mean that there is a fundamental assumption that what one wants to extract from the data is everything. However, in trying to solve the most microscopic version of a signal processing problem we may do a poor job (or no explicit job at all) of solving more macroscopic (and possibly more important) versions of the problem.

5. A significant amount of research in signal processing is technique-oriented rather than problem-oriented. That is, the processing algorithms are not based on an explicit specification of what information is to be extracted.

So what is our role in problems of this type? That role, as I see it, stems from the fact that we typically deal with precisely specified mathematical problems. This is important for several reasons which I will explicitly discuss in a moment. First, however, let me illustrate and motivate my points in the context of a specific signal processing problem.

A Model-Based Approach to Reconstruction from Line Integrals

The application I'll discuss is tomography, that is, the reconstruction of multidimensional functions from measurements of line integrals of these functions. The particular perspective and formulations I'll describe are presently being pursued in the doctoral research of David Rossi.

The general reconstruction problem can be stated as follows: let f(x) be the function to be determined, where x is an n-dimensional vector taking values in a specified bounded region of n-space. Our reconstruction of f(x) is to be based on possibly noisy measurements of integrals of f(x) along a (possibly infinite) set of lines $\{L_i\}$ through the specified region:

$$g(L_i) = \int_{L_i} f(x)\, ds \qquad (1)$$

For the sake of this discussion, I will concentrate completely on 2-D reconstruction (which is also the focus of our present work), although the ideas I'll discuss easily extend to higher dimensions.

Applications that require the use of reconstruction techniques are myriad. Easily the best known area of application is medical X-ray tomography. The fundamental principle behind Computerized Axial Tomography (CAT) scanners is that the attenuation of X-radiation through tissue is directly related to the line integral of the tissue's X-ray absorption density. A number of other medical applications can also be found. For example, ultrasound measurements provide time-of-flight measurements, which can be directly interpreted as measurements of integrals of tissue refractive index. Also, there is the problem of emission tomography in which radionuclides are injected into the body, and the detection of emitted positrons and gamma rays form the basis for the reconstruction of the distribution of the energy source and the tissue absorption density.

There are also a wide variety of non-medical applications. These include problems in electron microscopy, geophysics, radio astronomy, meteorology, nondestructive testing, target shape estimation, and oceanography. One interesting example in oceanography is based on the fact that the velocity of sound in water is temperature-dependent. Consequently, time-of-flight measurements, obtained by setting off a number of charges and recording the outputs of a set of acoustic receivers, can be used to obtain both sound velocity and underwater temperature profiles. Such profiles are of importance in military applications and also are potentially of great value in mapping and tracking large water masses such as Gulf Stream cold-core rings, which consists of a donut of warmer water (spun off of the Gulf Stream) with a cold center.

To begin our discussion of the analysis of reconstruction problems, consider the geometry of the problem as illustrated in Figure 5. Assume that $f(x)$ has its support inside the circle of radius T. Each line through this region is uniquely parameterized by two parameters (t,θ) which specify the polar coordinates of the vector from the origin perpendicular to the particular line. The integral of $f(x)$ along $\ell(t,\theta)$ is denoted by $g(t,\theta)$, which is known as the Radon transform of $f(x)$ in honor of J. Radon who in 1917 solved the problem of inverting the integral equation corresponding to the reconstruction problem when noise-free measurements of $g(t,\theta)$ are available for all values of t and θ.

To illustrate the geometry of Radon transforms consider the example depicted in Figure 6. Here $f(x)$ is zero except for x in the square region indicated in the figure. In this region it takes on a constant, nonzero value. Also indicated in the figure is $g(t,\theta)$ as a function of t for a fixed value of θ. In this case $g(t,\theta)$ is proportional to the chord length of intersection of $\ell(t,\theta)$ and the square. Also, note that the centroid of $g(t,\theta)$ as a function of t is $r \cos(\theta-\phi)$, which is the projection of the centroid of the square. These facts can also be seen in Figure 7, where $g(t,\theta)$ for this example can be seen as a "sinusoidal mountain range", where the cross-section of $g(t,\theta)$ depends on the angle θ and where the centroid of these cross-sections traces out a perfect sinusoid as a function of θ. Note that if $g(t,\theta)$ were displayed as an image with differing heights in Figure 7 transformed to differing intensities, the picture would simply consist of a single sinusoidal strip.

Now let us turn to the question of reconstruction. As mentioned earlier, the exact reconstruction of $f(x)$ from complete, noise-free measurement of $g(t,\theta)$ was first considered by Radon. More recent investigations have determined other forms for this inverse which have led to useful algorithms in practice. Specifically, one ideal reconstruction method relates the Fourier transform of $g(t,\theta)$(as a function of t) to the 2-D Fourier transform of $f(x)$. In addition, there is an alternate form known as a <u>convolution back-projection</u> algorithm. Specifically, it can be shown that

$$f(x) = \int_0^{\pi} \int_{-\infty}^{\infty} g(t,\theta) v(t-\tau(x,\theta)) \, dt \, d\theta \tag{2}$$

where

$$\tau(x,\theta) = x_1 \cos \theta + x_2 \sin \theta \tag{3}$$

and where the Fourier transform of $v(t)$ is $|\omega|$. The interpretation of this formula is illustrated in Figure 8. Note that the value of g at some point (t,θ) clearly provides information about the value of f at points x along the line $\ell(t,\theta)$. Intuitively, then what one might do is take each value of $g(t,\theta)$ and back-project it along the line $\ell(t,\theta)$. That is, we assign the value of $g(t,\theta)$ to every point along $\ell(t,\theta)$. The superposition of these back-projections then sould resemble the original function in that points of large f-value will have a significant number of large back-projected components. Note, however, that this approach leads to a smearing of $f(x)$ along lines (one obtains so-called "star patterns" if $f(x)$ has an isolated bright spot). Thus before back-projection one preprocesses each slice of $g(t,\theta)$ viewed as a function of t for fixed θ, by performing what amounts to bandlimited differentiation. This convolution operation effectively counteracts the smearing of back projection. Mathematically, one calculates $s(t,\theta) = g(t,\theta)*v(t)$ for each θ and then performs the back projection on $s(t,\theta)$ as depicted in Figure 8. Note that for a fixed x what this amounts to is integrating $s(t,\theta)$ along the cosinusoidal Radon space trajectory corresponding to lines through the point x.

In practice, of course, one has only a finite number of line integral measurements. Consequently, practical convolution back-projection algorithms involve sums rather than the integral in (2), where the discrete convolution kernal is chosen based on noise-resolution tradeoffs. Also an alternative set of algorithms results from a reformulation of the problem as an intrinsically discrete one. Specifically, a pixel representation of $f(x)$ is assumed, i.e., the support of $f(x)$ is broken up into small square cells over which $f(x)$ is assumed to be constant. With this assumption each

$g(t,\theta)$ is a weighted sum of the discrete set of values of $f(x)$, where the weights depend upon the chord length of intersection of $\ell(t,\theta)$ with each pixel. If we let f denote the vector whose components are the pixel values of $f(x)$, and if g denotes the vector of line integral measurements, we obtain a linear relation of the form

$$g = Hf \quad . \tag{4}$$

Thus the reconstruction problem has been posed as a problem of inverting this equation to determine f given g.

With this introduction, we can now make some observations:

(a) The assumptions of $f(x)$ that underly the formulations described so far are quite modest. Consequently the number of degrees of freedom is enormous. For example, in pixel-oriented representations (either using convolution back-projection or a method for solving (4) if one constructs an image consisting of a 256 x 256 array of pixels, there are more than 65,000 unknowns! To underline this point it is worth mentioning that typical X-ray measurement systems take on the order of 100-300 line integrals at each of 180 values of θ for a total of on the order of 18,000 - 54,000 measurements.

(b) In the most widely used methods, a priori information and the presence of noise are either not taken into account or are considered in the context of the solution technique rather than in the formulation. Consequently, they are dealt with in less than fundamental ways. For example, in convolution back-projection algorithms, a priori information enters only in terms of the desired resolution, and the presence of noise is taken into account in evaluating resolution-noise tradeoffs. In many of the deterministic approaches to solving (4), noise is really never taken into account directly, and a priori information only enters in indirect and heuristic ways aimed at allowing one to solve a set of equations with more unknowns than measurements. Finally, to my knowledge there are no existing methods that allow direct incorporation of a priori information of a structural nature (such as the knowledge of the presence of a bone or a cold-core ring).

(c) Because of the large number of degrees of freedom and the limited utilization of a priori information, obtaining an accurate reconstruction requires a large number of relatively high quality line integral measurements. This has at least three implications:

- CAT scanning involves subjecting a patient to a significant X-ray dose in order to obtain the required number of line integrals and the required signal-to-noise ratio in each measurement.

- The methods described so far are not of much use in situations in which it is essentially impossible to obtain the required data. For example, in problems such as tomographic reconstruction of the heart and the ocean temperature imaging problem described previously, only a limited number of viewing angles are possible (the ribs provide constraints in the former application, while prohibitive cost is one reason in the latter). It is worth noting that while the matrix H is well-conditioned if views are available at discrete angles over the entire range $0 \leq \theta \leq \pi$, this is not the case if view angle is restricted. The implication of this for noise performance is clear.

- There are some measurement systems (such as those involving acoustic energy) in which there are basic limitations in measurement signal-to-noise ratio due to the inaccuracy inherent in modeling the measurements as simple line integrals. Consequently, the utility of existing methods is questionable.

(d) Sometimes the ultimate goal of the processing is far more modest than the goal of estimating a 65,000-dimensional vector. For example, in medical applications one might want to detect the presence of a tumor or to outline tumors, bones, and organs. A classical problem in radiology is the detection of objects in noisy images. Detecting and localizing voids and faults in materials is another example.

Based on these observations, we have begun to look at a very different set of tomographic problem formulations. Specifically, we have in mind a model for the random field f(x), consisting of a background field on which are superimposed objects of differing intensity. For models of this type one can conceive of a hierarchy of problems:

Object detection: Here the problem is simply one of hypothesis testing. Is an object there or not?

Object localization: Given the presence of one or more objects, find out where each is located.

Object shape estimation: Given the location of an object, determine its boundary.

Object contouring: Given an object, map f(x) values within the object by finding contours of constant intensity.

Intuitively these are problems of increasing complexity. The first of these have modest goals, which hopefully can be matched by simplified formulations with reduced numbers of degrees of freedom whose solutions are robust to the simplifications that are made. Also, one might expect (and hope) that the use of simplified models and modest goals and the incorporation of structural a priori information would be rewarded by an increase in apparent signal-to-noise ratio. Our goal is to quantify ideas such as these by developing a thorough understanding of the assumptions underlying each problem and by examining the properties of the solutions to the problems we formulate.

Our motivation for formulating and studying problems of the types just introduce comes from a variety of applications. For example, consider the problem of low-dose, X-ray tomography for early detection of tumors. Here the ultimate goal is to make a binary decision: is an individual healthy or is there enough question about the presence of a tumor to warrant a full CAT scan? In this case the obvious questions that arise are: (1) how low can the dose be if a certain desired performance reliability is specified; and (2) how complex a model is needed to solve the problem accurately? As a second example, consider the oceanographic temperature and velocity mapping problem. Here the ultimate goal is a map, and the fundamental question is how accurately can we map when we have available a given quantity of data of a certain quality. The hierarchy of problems I've described seems well-matched to applications such as these. First of all, much of the fundamental information we have (and perhaps which we wish to extract) is object-oriented. Second, each successive problem requires inclusion of more detail which was neglected in the preceding problems. Furthermore, the problems as we've described them are nested, with successive problems using the solutions to preceding problems as starting points and including detail neglected by previous problems. This suggests an appealing algorithm structure in which detail is included only when we know where to focus it. At some level the data will be insufficient to determine the desired detail, and this should be quantifiable.

In addition to the two applications metioned so far, there appear to be a variety of others for which this perspective may be appropriate. For example, in the limited view heart reconstruction problem, we know a priori that the field being observed includes ventricles, atria, and, perhaps, an infarction. In the use of ultrasound for breast tumor detection, resolution may be poor, but the performance of an object detector may not be.

All of these wondrous things notwithstanding, all I've done so far is to suggest a perspective for formulating problems and a framework for analyzing the utility of different algorithms. We are presently at a very early stage in developing methods based on this perspective, but I would like to relate to you how our work is developing, as I think it exemplifies the systematic approach to model-based signal processing that I find personally appealing and satisfying. Simply put, what this approach has as its goal is the carving out of precise and tractable mathematical problems and the precise determination of what a particular formulation or algorithm is good for and what it is not good for. No miracle solutions claimed. Only bottom-line evaluations of utility and a framework for critiquing postulated formulations and solutions. I'm sure most will not find this approach revolutionary. I certainly don't. Most engineers do this sort of thing in some form or another. However, I do feel that this is an exciting perspective for signal processing, especially as more and more complex problems are considered.

As a start, we are examining just about the simplest problem involving fields with objects. Specifically, let K(c) denote a two-dimensional region (i.e., an object) of known shape and orientation, whose location is specified by the point

$$c = (c_1, c_2) = (r \cos\phi, r \sin\phi) \; . \tag{5}$$

We suppose that the field f(x) is given by

$$f(x) = f\chi_{K(c)}(x) \tag{6}$$

where $\chi_A(x)$ is the indicator function of the set A, and f is the constant intensity of the field on the object. That is, the field f(x) is assumed to have a constant known background intensity (which, without loss of generality, we take to be zero) and to have a different constant value on the object of known shape. An example of such a field was shown in Figure 6, in which K(c) is square of known size with centroid c. The only unknowns in the problem we've just formulated are the object location (i.e., c) and possibly f. As estimating the latter given the former is a simple problem, assume we know f as well. The problem then is to locate the object.

Suppose that what we observe are white noise-corrupted measurements of the Radon transform of f(x):[†]

$$y(t,\theta) = g(t,\theta) + w(t,\theta) \tag{7}$$

[†] I will phrase everything in terms of a white noise model and continuous-valued observations. It is also possible to consider the problem when y(t,θ) is a counting process with rate that is a function of g(t,θ). Such a model is appropriate for very low-dose X-ray problems where each count counts.

While we actually consider the case when $y(t,\theta)$ is observed for a discrete set of values of t and θ, it is convenient to demonstrate the basic ideas assuming complete measurements. Given such measurements we wish to compute the maximum likelihood (ML) or maximum a posteriori (MAP) estimate of c. To be specific, I'll stick to the former. In this case, we have a standard parameter estimation problem. The ML estimate of c is obtained as that value which maximizes the likelihood function L(c), which in this case is calculated according to the following: Let $g_K(t,\theta;c)$ be the Radon transform of (6). Then L(c) is given by

$$L(c) = \int_{-T}^{T} y(t,\theta)\ g_K(t,\theta;c)\ dt\ d\theta \tag{8}$$

Note, however from the geometry of the problem that

$$g_K(t,\theta;c) = g_K(t-\tau(c,\theta),\theta;0) \tag{9}$$

where τ is defined in (3). Thus

$$L(c) = \int_0^{\pi} \int_{-T}^{T} y(t,\theta) g_K(t - \tau(c,\theta),\theta;0)\ dtd\theta. \tag{10}$$

Comparing (2) and (10) we have that the calculation of L(c) is <u>precisely</u> a convolution back-projection operation, where the convolving function (which here is θ-dependent unless K is a circle) is the Radon transform of the object (located at the origin) for which we are searching.

As an aside note that one obvious way to locate or detect an object is to perform the usual convolution back-projection as in (2) and then apply some object localization or detection algorithm on the resulting image. Since the image typically has more points (65,000) than there are measurements

(20,000) it is highly unlikely that information is lost in this process. However, one must still find the object. Furthermore, not only have we increased the number of variables, but we've also lost the whiteness of the corrupting noise on each variable. On the other hand, if we use the specific convolution back-projection operation of (10), the localization (or detection problem) is trivial -- just find the largest value. Said another way, by formulating a specific problem we have found the convolving function g_K that is precisely matched to the specified task.

Let me close my discussion of model-based signal processing with several comments. Clearly the single object problem I've described is an exceedingly idealized one. Even if we are primarily interested in detecting and locating a single object (e.g., a tumor), there may be other objects (bones, organs) present. Furthermore the shapes, sizes, and orientations of the objects aren't known precisely, and the background and object intensities won't be constant or known. The key point is, however, that by examining the assumptions we've made, which are pinpointed by having explicit problem formulations, we can play devil's advocate with each problem and with algorithms for its solution. That is we can examine how the algorithm performs when some of the assumptions on which it is based are violated. In this way we can establish the limits on what particular problems/algorithms do and what they can't do.

What I have just described seems to me to be an appealing scientific approach to large scale signal processing problems which we are presently trying in the context of the object localization problem. That is, we are investigating the performance of our object localization algorithm when each of the assumptions we've made is violated, where our performance measure is the probability that the estimated centroid location is a point that is inside the actual object. It seems likely that the algorithm will be quite robust to actual object shape and intensity variations. If this proves to be the case it implies that one can use relatively simple models and algorithms for the detection and localization of objects. Once located, one can then consider the finer questions of the estimation of object goemetry and of intensity variations. Furthermore through this study we will also be able to establish what performance improvements would be possible if we directly considered more complex problems such as simultaneous localization and geometry estimation. In this way we will be able to provide an assessment of the range of situations in which the solution to the simpler problem would suffice.

I think it is also important to point out that this model-based approach to tomography does not consist solely of the application of existing theories and methodologies. Rather, in examining the inclusion of structural information we have uncovered several theoretical problems which are challenging but tractable and the understanding of which will shed light on the reconstruction problem in general and the potential of our approach in particular. Specifically, there are a variety of problems of a geometrical nature. First of all, detection and localization of objects is equivalent to the problem of detection and phase estimation for sinusoidal strips in (t,θ)-coordinates. (see Figure 7). From this perspective localization of several objects looks very similar to the problem of multi-target tracking. However, there is one interesting twist. Referring back to Figure 8, note that the plot of $s(t,\pi)$ is upside down. This is not an accident, as it isn't difficult to show that (t,θ)-space is a Möbius strip. This fact has some interesting consequences in terms of the design and convergence of iterative algorithms for estimating c. A more basic challenge arises when one addresses the problem of geometry estimation. There are a variety of ways in which one can introduce randomness into shape. The challenge is to understand each with regard to both the types of shapes that can be so modelled and the nature of algorithms resulting from a particular model.

More generally, tomography is one example of an inverse problem presently being examined in the context of signal processing. Many other examples exist -- problems in scattering, random media models, etc. -- and many of these might benefit significantly from a problem-oriented investigation. Perhaps the most large-scale ongoing investigation of this type by researchers in decision and control is that being run by Jerry Mendel at the University of Southern California on the topic of seismic signal processing. I think Jerry would agree with me that there is a great opportunity for significant contributions in a wide variety of signal processing problems.

As an aside, let me make one comment about artificial intelligence. Specifically, the concept of problem decomposition is one that is often used by researchers in artificial intelligence. The approach that I have described for solving tomographic reconstruction problems also involves problem decomposition. As I see it, there is a natural marriage here. The perspective of AI is aimed directly at attacking and breaking down problems of enormous complexity into smaller problems. On the other hand, the perspective in control, estimation, decision, and system theory is to solve very precisely specified (and usually small) problems and to provide the means for quantitative evaluation of performance for these solutions. Although I certainly have not made explicit use of AI concepts, the spirit of what I've discussed suggests a view I hold of where there might be significant payoff from a dialogue among researchers in AI and in decision and control. Specifically, in order to break down a complex problem into manageable and solvable pieces, one must have a feeling for what one can solve.

Finally, it is obvious that things are not as clear-cut as I've made them seem. Technique-oriented researchers certainly solve problems. However, this approach to research certainly does tend to be much more bottom-up, where solutions to mathematical problems are used as techniques and aren't tied explicitly to the fundamental problem under investigation. (I would cite spectral estimation techniques as one class that is often used in this way). There are clear advantages to this approach. First, the methods that are used are typically well-understood (thanks to numerous successful applications), and the intuition associated with them is of great value to the signal processor in interpreting his results. Second, by not being married to specific types of models, these techniques may be more broadly applicable than model-based techniques and consequently allow one to get one's hands on the data more readily.

On the other hand, there are some disadvantages, or, more precisely, some limitations to this philosophy. The first is that there is no systematic framework for the incorporation of a priori information. Second, if something goes wrong, there is no systematic basis for determining precisely what has gone wrong and why. Third, and most importantly, the technique-oriented philosophy by its very nature focuses more on the techniques than on the problem. Consequently there is a danger that techniques that really are not appropriate will be forced on a problem. In some cases what this does is simply transform the fundamental problem to one of equal (and sometimes greater!) complexity. In addition, another potential danger is that it is very easy in a technique-oriented approach to avoid essentially completely any fundamental discussion of what the problem is. In large, complex signal processing problems it is often true that the largest and most important problem is in figuring out what the problem is!

In contrast to the technique-oriented viewpoint, the problem and model-oriented philosophy uses far more of a top-down perspective. There exist some clear limitations and distinct dangers with this approach if attempts are made to force physical problems into mathematical formulations with which we feel comfortable but which are totally inappropriate or if we naively place too much faith in solutions to mathematical problems without thoroughly critiquing the assumptions underlying our formulations. The point is that one can't use the problem-oriented approach in too top-down a manner. Said another way, I don't expect to see the field of signal processing revolutionized by papers with titles like "The Kalman Filtering Solution to the Weather Prediction Problem" (work of the type suggested by this title is really technique- rather than problem-oriented). While the specific approaches we have developed in other contexts will not doubt be of value, the real key to our role in signal processing is more conceptual in nature.

Specifically, a background in the development of mathematical methodologies provides one with insight as to what problems are tractable and what ones might be. This is valuable in uncovering useful formulations to complex problems. Furthermore the use of mathematical problems allows one to incorporate a priori information and, more importantly to engage in the type of scientific investigation I have described previously. That is, the use of precisely stated problems provides the basis in an application for evaluating models, mathematical problems, and solutions; for pinpointing assumptions and the limitations they imply; and for finding a model and problem formulation that is compatible with the goals of the processing and with the available quantity and quality of data. In addition, the process of demanding precision of thought by focussing on the construction of mathematical problems is in itself exceedingly valuable, as this process forces us to organize, analyze, and question our understanding of the phenomenon under investigation and of what we want to extract from it. Finally, this process allows researchers with very different perspectives to interact fruitfully by providing a common focus and point of reference -- a problem or the construction of a problem. It is my feeling that out of such interactions can come innovative approaches that are not likely to emerge as quickly if interaction is not fostered.

Some Questions

Now on to some questions. As I indicated earlier in this talk, what I am interested in doing is in stirring up some debate and discussion among researchers in our field. I have tossed out one research direction--model-based signal processing--which I feel is worthy of discussion and consideration, and I would personally welcome hearing of other promising research directions that people feel are important topics for the present and future. I would also like to see discussions of existing areas of reserach, as such examinations help keep research topics vital and also acquaint non-specialists with the essential concerns, ideas, perspectives, and burning issues that define an area.

As rolling friction is less than starting friction, I would like to pose a few questions that hopefully will get things rolling and spark discussion and the formulation of additional questions. I sincerely hope that my questions are accepted in the spirit in which they're given. First of all, they are not meant as judgments on particular aspects of the field but are intended only to stimulate debate and not defensive responses. I certainly have my own biases, and these definitely color the phrasing of my questions, but this really is beside the point, as my personal opinions here are secondary. Furthermore, it is my personal belief that the field of decision and control is thriving, and I can't think of a better time to ask questions and to engage in debate than when things are going well. Finally, I believe that an absolute prerequisite for fruitful and open discussion is the ability to laugh.

I have not asked questions about every major research direction due to my own ignorance, a lack of time to cover everything, or my inability to think of something clever to say. Consequently, I'd like to apologize in advance to those researchers who are not offended by my questions. There was no conscious effort made to avoid any particular research topic, so I hope you don't take it personally.

My first question is on the topic of large scale systems. In my opinion there is no doubt that this is one of the most important research directions before us. However, I know that I have a great deal of difficulty in figuring out how to say "large" in a problem formulation and how to develop manageable ways to deal with truly large systems. My question, as illustrated in Figure 9, is motivated by these difficulties. Specifically, a substantial portion of large scale systems research seems to deal with trying to force problems into the framework of tried and true methods of estimation, control, and analysis. Clearly in trying to understand a new type of problem an important step is that of trying to cast the problem in a familiar way, but in this case, this step would seem to be far from the final one.

My question, then, has two sides. First, what if any truly significant breakthroughs in large-scale systems have resulted from using solely the same methods and kinds of problem formulations that have been used for "small-scale systems"? Second, what are the truly new and innovative approaches to formulating large-scale problems and to thinking about large-scale systems?

As I view myself as very much a novice in this research area, I would personally learn a great deal from listening to the opinions and thoughts of leaders in the field. However, as this is <u>my</u> lecture, I will offer a few thoughts about results and perspectives that have intrigued me and that I find consistent in spirit with the approach I described for formulating and solving problems in signal processing. Specifically, the lines of research that have attracted me are those which involve looking for natural ways in which to break large problems and/or large systems into smaller pieces in such a manner that the solutions to the pieces can be molded together into a workable solution for the overall system.

For the sake of brevity, let me focus on one such line of research which is quite close to the types of problems with which many of us are comfortable, in that the starting point is single, precisely specified, familiar-looking problem. The novel feature of this formulation is that the structure of the problem naturally suggests a procedure for decomposing the problem, for putting the smaller solutions back together, and for assessing the performance of the overall design. I'm speaking of research, associated with names such as Kokotovic, O'Malley, Delebesque, Quadrat, Haddad, and Sandell, which has dealt with using the time scale separation inherent in large systems to reduce large problems to sets of far simpler ones at different time scales.

Several comments about this. First of all, people who've had to design or operate large systems have been doing this heuristically for years based on extensive experience with specific systems. At one level what the relatively recent research in this area has done is attempt to provide a systematic and quantitative foundation for decompositions of this type. Second, in my opinion the truly critical part of this work is that once the problem is stated, it is relatively easy to guess the solution. To be sure analysis must be done to assess the solution and show its asymptotic properties, but what I find really appealing is that the structure of the problem is used in a very fundamental way. Third, what I have just termed a great strength of this research has in fact been viewed by some as its great weakness. "Well if you know the structure of the system and problem so well that the solution pops out, it can't be a really large system. Anyway, most large-scale systems aren't given to you in this form." There are several responses to those assertions. Specifically, what this research has shown is the payoff to be obtained if one can find such structure in large systems. What this naturally suggests is a direction for theoretical research aimed at exposing and exploiting time scale structure for systems in which this structure is not transparent. Such research is being explored by a number of people at institutions around the world. In our work we've found that this endeavor is forcing us to ask some basic questions concerning the types of problems we really want to formulate (and in particular concerning the ways in which we evaluate performance), and this is proving to be one of the most exciting and gratifying aspects of our research.

A second response to the assertions stated previously is that in most large problems one is not working from a state of maximal ignorance. Typically people intimately involved with a large system know a great deal about the nature of the system and of the desired objectives. In such cases the concept of multiple time scales analysis plays the role of one potentially useful methodology which we can attempt to use together with the knowledge we have to structure a model and problem formulation that lead to a solution which is tractable and which meets the desired objectives. It seems to me that providing the designer or operator of a large-scale system with several such classes of techniques, each of which has been thoroughly and precisely analyzed in terms of its performance and its limitations (as determined by the assumptions on which it is based). is an admirable goal for theoretical research in large-scale systems.

The second subject on which I'd like to ask a question is one in which most of us have at least dabbled. The illustration for this question (Figure 10) is Susanna and my one attempt at allegory. My question is: Will linear system theory go on forever? Clearly linear systems will be with us forever, and it often makes eminent sense to examine new systems concepts, such as large scale systems methods, in the context of linear systems, thus avoiding the distracting complexities of nonlinearities. Furthermore, linear systems are extraordinarily rich, and there are enormously beautiful ideas and relationships to be extracted. But where is linear system theory going? Obviously a great deal of the energy in linear systems research is devoted to problems motivated primarily by their intellectual content. In addition, there is a substantial body of research that is aimed at issues raised by the needs of applications, such as in robust multivariable control. However, in either case I think that it is important to take a deep, hard look at where the research is heading in order to provide a clear statement of the important challenges and novel questions to be faced.

Such a critical evaluation is especially important in this field, as we are talking about the research area which regularly produces the most papers submitted to the IEEE Transactions on Automatic Control. As a relative outsider I am frankly overwhelmed by the volume of work and find it very difficult to gain an appreciation for what's important and significant and what is not. Consequently, I would really like to hear answers to some of the questions I haven't the knowledge or insight to answer. Where is the need for new and improved enormous Riccati equation solvers and for faster algorithms? Are there still important problems involving realization theory? What else is left to be done in the albegraic and geometric theories of linear systems? What really needs to be done to make multivariable control design methods useful and where is the present need for them? Related to this, let me refer you to the letter of Dr. J.R. Leigh which appeared in the June 1981 Control Systems Magazine in which he asks why advanced process control systems are still based on PID concepts and not on more recent theoretical methods. I for one would benefit from reading a response to this letter.

My third question is on robotics. At present this is certainly a topic that's creating quite a stir. Everybody seems to be talking about it or has an interest in it. In fact, to overstate a point, let me refer you to Figure 11. Now I know something of the feeling Dustin Hoffman must have had in The Graduate, only here the word is "robotics", rather than "plastics". My question then is: What should we be doing in robotics? Any of use can dream up a great many problems that sort of sound like they would be of importance in robotics. Doing this, however, leaves me with an uneasy feeling that such problems might not be the right ones, and the reason for this is that I personally don't have a clear picture of what the context is. What are the precise needs in automating industry? How flexible (i.e., intelligent, perceptive, etc.) must we make robots? What is the most cost-effective way in which to use robots? Here I have in mind that one might design an incredibly sophisticated robot at great cost (in dollars and in time and creative energy) when a simpler robot might do if its job were slightly simplified. For example, is it always necessary to have robots which can pick up parts that are placed in arbitrary orientations, or in some applications might one design a system so that parts arrive with only small deviations from a given orientation? I have come to share the view which I first heard put forth by my colleague Stan Gershwin that a systems-oriented view is of essential importance in determining the real needs for intelligent automation in any given situation. My guess is that the people who are deeply involved in robot research have done some of this type of thinking and have uncovered specific needs which have become their foci. Again at meetings such as the CDC what we tend to see are the results of these focused efforts rather than the thought process that motivated them. Perhaps it would be useful to have an open, interdisciplinary discussion among researchers aimed at exploring the systems view of the role of robots. Not only would this expose the perspective and context behind ongoing robotics research in the field of decision and control, but my guess is that it would also show that constructing such a systems view is a challenging research topic in itself.

My fourth question and the illustration for it (Figure 12) represent the exaggeration of an impression. The question is: What is the adaptive control problem or what are the adaptive control problems? It is evident that adaptive systems are of great importance in a wide variety of applications. It is also evident (at least to me) that there is an unbelievably confusing array of methods, partial results, successes, failures, ideas, beliefs, etc. What I have trouble finding are the themes, that is the precise problems and perspectives that are driving this field. Knowing these I might be able to find the key that would provide an orderly way of viewing all of the different adaptive schemes that abound. Without this, I feel like I'm walking through the Grand Bazaar in Istanbul. Lots of intriguing things, but it's awfully confusing.

Said another way, it is my opinion that adaptive control is the most technique-oriented research direction in decision and control. Perhaps this is the best way to approach adaptive problems and perhaps the fundamental ideas are absolutely clear to many who work in the field. However, I for one get uneasy when I see the forms of algorithms dictated by the desire to use a particular tool (the positive real lemma) for proof of convergence. Maybe this is fundamental. If it is and the reasons are clear, I'd like to see them explosed. If this is an open question, I'd like to see it exposed as such. This specific point is just one example. There are numerous other issues and ideas associated with adaptive control that I feel would benefit from being brought into clearer focus so that one could discuss what is fundamental and what is technique.

My fifth and final question is on the topic of nonlinear filtering. There is no doubt that this is a very difficult area for theoretical research. Thus any results and insights are to be valued as significant breakthroughs as they add substantially to our limited understanding. On the other hand, if one looks at many applications in which nonlinear estimation problems arise, one finds that somehow satisfactory solutions are constructed. After all, at one level all that's involved is Bayes' rule and the Chapman-Kolmogorov equation, and if one is willing to invest the computational effort, one can approximate these operations to any desired level of accuracy. Also, of course, there are ad hoc procedures like the extended Kalman filter which sometimes work satisfactorily. Not very pretty or intellectually satisfying, but they get the job done. Because of this, I find myself at times wondering (Figure 13) if what nonlinear filtering theory is doing amounts to making better and better tie clasps and belt buckles," but still the only set of clothes the Emperor has is his old pair of overalls. Thus my question is: What are the likely ways in which present research in nonlinear filtering will impact the design and analysis of nonlinear filters?

It is my personal belief that some important things will come out of present research directions. For example, what is developing is a way in which to determine and categorize the complexity of nonlinear filtering problems. That is, an understanding is emerging of why nonlinear filtering is hard. Said another way, the Emperor may still have to wear his overalls, but now he'll know why. As for the development of new approaches to the design of nonlinear filters, I think that the picture is far more uncertain. There are now a number of examples of filtering problems for which the optimal filter is finite-dimensional. In my opinion these examples are of importance primarily in developing an understanding of the complexity of nonlinear filtering and have no direct impact on filter design. On the other hand, what may be more promising are the several approaches to series and asymptotic expansions of nonlinear filtering solutions that are being pursued. Some of these are quite generally applicable, and others have been developed based explicitly or implicitly on assumptions of small noise and/or small nonlinearities. There is a third possibility here as well, dealing essentially with systems with *very* strong nonlinearities. This is an area in which several of us at M.I.T. are presently involved, and I would like to say a word about it, as it is consistent with one of the themes of this talk: if a problem is intractable, it is worth spending effort trying to change the problem. In the context of the problems we are considering it appears that one potentially useful way of changing the nonlinear filtering problem is to change the criterion. Specifically, our idea is that the problem of finding an estimator that is good (i.e., optimal or close to optimal) *at every point in time* (which is what criteria such as minimum variance imply) is hard. Furthermore, any such estimator will most likely be extremely complex. On the other hand, if one is willing to settle for an estimator that is good *most of the time* or *on the average*, it may be possible to find far simpler algorithms. It is far too early to say exactly what will come from this effort, but I feel that the idea of the effort in itself is important, as it puts forth the thought that maybe there are alternatives to the usual nonlinear filtering formulation which are more appropriate and more tractable in some cases.

This concludes what I have to say. Let me repeat that I am extremely honored to have been given this opportunity to relate some of my thoughts to the decision and control community. I'd also like to reiterate my belief that it is extremely important and worthwhile to engage in a fundamental, critical examination of what we do and why we do it. It is fine to lead by example, i.e., by publishing solutions to problems we've solved, but we also have the responsibility to lead by questioning, explaining, and providing perspective. Furthermore, I think that some of the likely things (in addition to new problems) that can come out of such examinations are new ways of looking at old problems which can breathe even more life into fields of research that are already pretty lively.