Signal Detection in a Nonstationary Environment Reformulated as an Adaptive Pattern Classification Problem

SIMON HAYKIN, FELLOW, IEEE, AND DAVID J. THOMSON, FELLOW, IEEE

The primary purpose of this paper is the improved detection of a nonstationary target signal embedded in a nonstationary background. Accordingly, the first part of the paper is devoted to a detailed exposition of how to deal with the issue of nonstationarity. The material presented here starts with Loève's probabilistic theory of nonstationary processes. From this principled discussion, three important tools emerge: the dynamic spectrum, the Wigner–Ville distribution as an instantaneous estimate of the dynamic spectrum, and the Loève spectrum. Procedures for the estimation of these spectra are described, and their applications are demonstrated using real-life radar data.

Time, an essential dimension of learning, appears explicitly in the dynamic spectrum and Wigner–Ville distribution and implicitly in the Loève spectrum. In each case, the one-dimensional time series is transformed into a two-dimensional image where the presence of nonstationarity is displayed in a more visible manner than it is in the original time series. This transformation sets the stage for reformulating the signal detection problem as an adaptive pattern classification problem, whereby we are able to exploit the learning property of neural networks. Thus, in the second part of the paper we describe a novel learning strategy for distinguishing between the different classes of received signals, such as: 1) there is no target signal present in the received signal; 2) the target signal is weak; and 3) the target signal is strong.

In the third part of the paper we present a case study based on real-life radar data. The case study demonstrates that the adaptive approach described in the paper is indeed superior to the classical approach for signal detection in a nonstationary background.

I. INTRODUCTION

The detection of a target signal in a background of noise is basic to signal processing. The term "noise" is used here in a generic sense; its exact description depends on the application of interest. The need for signal detection arises in many diverse fields, as summarized in Table 1. The detection problems summarized in this table may be grouped under two headings, depending on how the raw data originate:

1) *Time Series*—radar, sonar, digital communications, and scientific data;
2) *Images*—magnetic resonance images, positron emission tomographs, mammograms, X-ray photographs, nondestructive testing of metals, detection of explosives in carry-on bags, and detection of land mines.

In this paper, we focus attention on the first class of signal detection problems. Nevertheless, many of the ideas described herein also apply to the second class.

The classical approach to the design of a signal detection system is to optimize a time-invariant (i.e., fixed) receiver structure for a known class of signals with unknown parameters that are corrupted by noise of known statistics [1]–[3]. The parameter used in the design is the likelihood ratio of the received signal. When, however, the noise is nonstationary but is of known form, the receiver must take on a time-varying structure, making its design more difficult [4]. The design becomes even more difficult when the statistics of the noise are unknown. Our interest in this paper lies in the class of signal detection problems described as that of detecting a nonstationary target signal in a nonstationary environment of unknown statistics. This is the most difficult class of signal detection problems. For example, in radar, sonar, and mobile communications systems, nonstationarity of the received signal arises due to variations in environmental conditions. Although it may be feasible to view the received signal in such systems as quasi-stationary in the very short term, and possibly stationary in long-term averages, at the intermediate time-scales useful for engineering applications the received signal can be significantly nonstationary. In radio systems, for example, it is usually a reasonable assumption that the continuum (as opposed to impulsive) noise may be approximately stationary for ms to s, and also if it is averaged over months. The solar component of such noise,

Manuscript received November 1, 1997; revised April 17, 1998.

S. Haykin is the Communications Research Laboratory, McMaster University, Hamilton, Ont. L8S 4K1 Canada.

D. J. Thompson is with Lucent Technologies, Murray Hill, NJ 07974 USA.

Publisher Item Identifier S 0018-9219(98)08070-0.

Table 1 Areas of Application Where Signal Detection Arises

Area	Target	Main source of interference
1. Radar	Echo from Aircraft, ship, etc.	Clutter due to radar backscatter from unwanted objects, noise
2. Electronic-counter-counter measures (ECCM)	False radar transmission	Radar clutter, noise
3. Impulse (ground-penetrating) radar	Buried objects	Clutter due to reflections from the ground, noise
4. Sonar	Echo from submarine	Reverberation due to reflections from the ocean body, noise
5. Digital Communications Multiuser detection in code-division multiple access systems	Symbol 1 or Symbol 0 Sequence of 1s and 0s	Intersymbol interference or multipath, noise Interference from other users, multipath, noise
6. Biomedical signal processing	Epileptic seizure in the EEG signal of an infant	Normal EEG signal
7. Magnetic resonance images (MRI)	Abnormal distribution of tissue	Neighboring tissues
8. Positron emission tomography (PET)	Regional measurement of metabolism	Normal metabolism in the neighboring tissues
9. Mamograms	Low-contrast anomalies in the tissue	Spatial and contrast resolutions
10. Chest X-ray photographs	Abnormal tissue	Very high dynamic range of spatial resolutions
11. Nondestructive testing of metals	Defects	Clutter due to reflection from main body of metallic object
12. Explosives detection: X-ray image	Explosives, firearms, knives	Presence of harmless objects in carry-on luggage
13. Detection of land mines	Echo from a mine buried in the ground	Reflections from the ground and other objects of no interest
14. Detection of a flaw in textured clothing material	Flaw in the texture	Normal texture

however, varies on both the 5-min scale of the solar p-modes [5], and also with the 26-day solar rotation and the 11-year solar cycle [6]. In the radar data used in this paper we find unexpected evidence for cyclostationary, or periodically correlated, behavior in the clutter.

To deal with the issue of unknown statistics of the additive noise, we may use a neural network (NN) to compute the likelihood ratio of the received signal by training it on different realizations of the received signal. Such an approach is described in [7]–[9]. The NN's described in those references belong to the class of focused time-lagged feedforward networks, focused in the sense that a short-term memory structure (used for dealing with time) is confined entirely to the input layer of the NN. In [10] it is shown that these networks are universal approximators of myopic nonlinear dynamic systems. Unfortunately, however, their use is limited to stationary processes. To deal with a nonstationary process using an NN approach, the implicit effect of time has to be distributed inside the synaptic structure of the NN as described in [11], or else a recurrent NN has to do the learning in a dynamic fashion [12].

In this paper, we take a different approach for dealing with signal detection in a nonstationary environment. We proceed by first computing two-dimensional (2-D) image representations of the received signal that account for the nonstationary nature of the signal. In so doing, we account for time in an explicit sense, thereby transforming the detection problem into an adaptive pattern classification problem that lends itself to learning [13]. Such an approach to adaptive signal detection is described in [14]–[17]. This approach is philosophically distinct from classical detection procedures: the parsimonious, but frequently overly simplified models of signal and noise are replaced by a highly redundant and overcomplete representation of the received signal.

The main body of the paper is organized as follows. Section II discusses the nonstationary behavior of a signal and the related issue of time-frequency analysis in general terms. Section III presents a theoretical background for dealing with nonstationary signals. This discussion leads naturally to an overview of procedures for estimating the spectrum of a nonstationary signal in Section IV. In Section V we present some results on the time-frequency analysis of real-life clutter data as an illustrative example of the procedures described. Section VI describes a modular learning strategy for the detection of a target signal embedded in a nonstationary background. Section VII presents highlights of a case study that builds on this detection philosophy. Section VIII discusses the issue of cost functions for supervised training of the pattern classifiers in the adaptive receiver. The paper concludes with some final remarks in Section IX.

II. An Overview of Nonstationary Behavior and Time-Frequency Analysis

The statistical analysis of nonstationary signals has had a rather mixed history. Although the general second-order theory was published during 1946 by Loève [18], [19], it has not been applied nearly as extensively as the theory of stationary processes published only slightly previously by Wiener and Kolmogorov. There were, at least, four distinct reasons for this neglect.

1) Loève's theory was probabilistic, not statistical, and there does not appear to have been successful attempts to find a statistical version of the theory until some time later.
2) At that time of publications, the mathematical training of most engineers and physicists in signals and random processes was minimal and, recalling that even Wiener's delightful book was referred to as "The Yellow Peril," it is easy to imagine the reception that a general nonstationary theory would have received.
3) Even if the theory had been commonly understood at the time and good statistical estimation procedures had been available, the computational burden would probably have been overwhelming. This was the era when Blackman–Tukey estimates of the stationary spectrum were developed, not because they were great estimates but, primarily because they were computationally more efficient than other forms.
4) Finally, it cannot be denied that the general theory was significantly harder to grasp than that for stationary processes.

Nonetheless, it was realized that many, perhaps most, of the signals being worked with were nonstationary and, starting with the available tools, i.e., the ability to estimate the spectrum of a stationary signal, the spectrogram was developed. The idea was that, if the process is not "too" nonstationary, then for a relatively short time block a "quasi-stationary" approximation can be used, so that for the length of the block the spectrum can be approximated by its average. It was also recognized that a major drawback of the spectrogram is that the block lengths and offset between blocks are arbitrary. Thus, although speech, underwater sound, radar, and similar communities have much empirical experience to guide such choices, little can be done with a new, possibly unique, data series except "cut and try" methods. Consequently, it is common to regard the spectrogram as a heuristic or *ad hoc* method.

To account for the nonstationary behavior of a signal, we have to include time (implicitly or explicitly) in a description of the received signal. Given the desirability of working in the frequency domain for well established reasons, we may include the effect of time by adopting a time-frequency description of the signal. During the last 20 years many papers have been published on various estimates of time-frequency distributions; see, for example, [20] and the references therein. In most of this work, the signal is assumed to be deterministic. In addition, many of the proposed estimators are constrained to match time and frequency marginal density conditions. If $D(t, f)$ is a time-frequency distribution of a signal $x(t)$, it is required that the time marginal satisfy the condition

$$\int_{-\infty}^{\infty} D(t, f)\, df = |x(t)|^2$$

and, similarly, if $y(f)$ is the Fourier transform of $x(t)$, the frequency marginal density must satisfy the condition

$$\int_{-\infty}^{\infty} D(t, f)\, dt = |y(f)|^2$$

where t denotes continuous time and f denotes frequency. Given the large differences observed between waveforms collected on sensors spaced short distances apart (see, for example, [21]), the time marginal requirement is a rather strange assumption. Worse, the frequency marginal distribution is, except for a factor of $1/N$, just the periodogram of the signal. Since it has been known since at least the 1930's that the periodogram is badly biased and inconsistent,[1] and we have personally experienced engineering data [22] where the periodogram was wrong by more than a factor of 10^{10} over most of the frequency range, imposition of such a constraint must be viewed skeptically as well. Thus we do

[1] A biased estimate is one where the expected value of the estimate differs from the value of the quantity being estimated. An inconsistent estimate is one where the variance of the estimate does not decrease with sample size. The periodogram is an unstable, wrong, answer.

not consider matching marginal distributions, as commonly defined, to be important.

Similarly, several estimates have been proposed that attempt to reduce the cross terms[2] in the Wigner–Ville distribution (WVD) by using the analytic signal instead of the original data. However, as the analytic signal is commonly derived by Fourier transforming the data, discarding the negative frequency components, and taking the inverse Fourier transform, the frequency-domain bias of the analytic signal is dominated by the periodogram bias. The opinion has been expressed that these concerns apply only in a near-pathological data set and that the sidelobe performance of the Slepian sequences is rarely needed. We consider this opinion ill advised as we rarely know in advance what sidelobe performance is needed. (Slepian sequences are defined in Section III-B.) As an example, the dynamic range of the spectrum for the radar data used in this paper exceeds 10^4, so an estimate constrained to match periodogram marginals could easily be in error by an order of magnitude over most of the frequency domain.

This being said, the WVD used in the following is computed by the standard, basic form. We have several reasons for leaving the WVD unaltered.

1) As we describe it below, the expected value of the WVD is just a coordinate rotation of the Loève spectrum. It is not, however, a particularly good statistical estimate.
2) The basic WVD is a sufficient statistic in that it can be inverted to recover the original data to within a phase constant [23]. Thus, although it is not an attractive estimate from a statistical viewpoint, its completeness properties allow it to be effective in our application (i.e., signal detection).
3) The data used here are complex valued, so there is no need to estimate the analytic signal.
4) The cross terms are visually distinctive and so may be a significant help in recognizing that more than one component is present in the received signal.

In the final analysis, whether we adopt the stochastic or deterministic approach to time-frequency analysis for representing the nonstationary behavior of a signal depends on details of the problem of interest. It is easy to imagine problems where one or the other of these two approaches would be preferable, but there are other problems, such as the radar data used in our examples, where a good case can be made for both viewpoints.

III. Theoretical Background

Suppose we are given data consisting of a single finite realization of N contiguous samples of a discrete-time process $x(t)$ for $t = 0, \ldots, N-1$; henceforth, t denotes discrete time. We assume that the process is harmonizable [19] so that it has the Cramér, or spectral, representation

$$x(t) = \int_{-1/2}^{1/2} e^{j2\pi\nu t}\, dX(\nu) \tag{1}$$

where $dX(\nu)$ is the increment process. In this paper, we also assume that the process has zero mean, that is, $\boldsymbol{E}\{dX(\nu)\} = 0$, and, correspondingly, $\boldsymbol{E}\{x(t)\} = 0$. (Note that this is not the same as assuming that an average has been subtracted from the data.) As parameters of interest, consider then the covariance function

$$\begin{aligned}\Gamma_L(t_1, t_2) &= \boldsymbol{E}\{x(t_1)x^*(t_2)\} \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{j2\pi(t_1 f_1 - t_2 f_2)} \gamma_L(f_1, f_2)\, df_1\, df_2\end{aligned} \tag{2}$$

and the generalized spectral density

$$\gamma_L(f_1, f_2)\, df_1\, df_2 = \boldsymbol{E}\{dX(f_1)\, dX^*(f_2)\} \tag{3}$$

where * indicates complex conjugate. Equation (3) describes the essential feature of nonstationary processes, namely, that there is correlation between different frequencies.

If the process is stationary, the covariance $\Gamma_L(t_1, t_2)$ depends (by definition only) on the time difference $t_1 - t_2$, and the Loève spectrum $\gamma_L(f_1, f_2)$ becomes $\delta(f_1 - f_2)S(f_1)$ where $S(f)$ is the ordinary power spectrum. Similarly, for a white nonstationary process, the covariance function becomes $\delta(t_1 - t_2)P(t_1)$ where $P(t)$ is the expected power at time t. Thus, as both the spectrum and covariance functions include delta function discontinuities in simple cases, neither should be expected to be "smooth," and continuity properties depend on direction in the (f_1, f_2) or (t_1, t_2) plane. These problems are more easily dealt with by rotating both the time and frequency coordinates of the generalized correlations (2) and spectral densities (3) respectively by 45°. In the time domain, we define the new coordinates to be a "center" t_0 and a delay τ, as shown in the following:

$$\begin{aligned} t_1 + t_2 &= 2t_0 \\ t_1 - t_2 &= \tau. \end{aligned} \tag{4}$$

Equivalently, we may write

$$\begin{aligned} t_1 &= t_0 + \tau/2 \\ t_2 &= t_0 - \tau/2. \end{aligned}$$

We denote the covariance function in the rotated coordinates by $\Gamma(\tau, \tau_0)$ and so write

$$\Gamma_L(t_1, t_2) = \Gamma\left(t_1 - t_2, \frac{t_1 + t_2}{2}\right) = \Gamma(\tau, t_0). \tag{5}$$

Similarly, we define new frequency coordinates f and g by writing

$$\begin{aligned} f_1 + f_2 &= 2f \\ f_1 - f_2 &= g. \end{aligned} \tag{6}$$

[2] Cross terms arise when the Wigner–Ville distribution is applied to the sum of two signals. The Wigner–Ville distribution of such a sum is not equal to the sum of the Wigner–Ville distributions of the two signals; the difference is accounted for by the cross-terms.

Equivalently, we may write

$$f_1 = f + g/2$$
$$f_2 = f - g/2.$$

Denote the rotated spectrum by

$$\gamma(g,f) = \gamma_L\Big(f + \frac{g}{2}, f - \frac{g}{2}\Big). \tag{7}$$

Substituting these definitions in (2) shows that the term $t_1 f_1 - t_2 f_2$ in the exponent of the Fourier transform becomes $(t_0 g + \tau f)$, and so

$$\Gamma(t_0,\tau) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{j2\pi(\tau f + t_0 g)}\gamma(g,f)\,df\,dg. \tag{8}$$

Because f is associated with the time difference τ, it corresponds to the ordinary frequency of stationary processes and we refer to it as the "ordinary" or "stationary" frequency. Similarly, because g is associated with the average time t_0, it describes the behavior of the spectrum over long time spans, and we refer to g as the "nonstationary" frequency. Now consider the continuity of γ as a function of f and g. On the line $g = 0$, the generalized spectral density γ is just the ordinary spectrum with the usual continuity (or lack thereof) conditions normally applying to stationary spectra. As a function of g, however, we expect to find a δ function discontinuity at $g = 0$ if, for no other reason, that almost all data contain some stationary additive noise. Consequently, smoothers in the (f,g) plane (or, equivalently, the (f_1,f_2) plane) should not be isotropic, but they require much higher resolution along the nonstationary frequency coordinate than along the ordinary frequency axis f.

A slightly less arbitrary way of handling the g coordinate is to Fourier transform $\gamma(g,f)$ with respect to the "nonstationary" frequency g and define

$$D(t_0,f) = \int_{-\infty}^{\infty} e^{j2\pi t_0 g}\gamma(f,g)\,dg$$

as the theoretical "dynamic spectrum" of the process. The motivation is to transform the very rapid variation expected around $g = 0$ into a slowly varying function of t_0 while leaving the usual dependence on f. Because δ functions in frequency transform into a constant in time, in a stationary process $D(t_0,f)$ does not depend on t_0 and becomes $S(f)$. Writing D as

$$D(t_0,f) = \int_{-\infty}^{\infty} e^{j2\pi\tau f}\boldsymbol{E}\Big\{x\Big(t_0 + \frac{\tau}{2}\Big)x^*\Big(t_0 - \frac{\tau}{2}\Big)\Big\}\,d\tau \tag{9}$$

we see that the rotated Loève spectrum is the expected value of the WVD. This relation has been rediscovered several times (see, for example, [24]). Note carefully, however, that unlike the standard WVD, D is defined to be an expected value.

Stated in another way, the WVD is the instantaneous estimate of the dynamic spectrum $D(t_0,f)$ and therefore simpler to compute than the dynamic spectrum. Taking the complex conjugate of (9) shows that D is real, and, as the Fourier transform of a covariance, must be nonnegative definite (see [25]). In many ways, current discussions about positivity of the WVD and similar distributions are reminiscent of those occurring in papers in the 1945–1970 era on whether the normalization $1/N$ or $1/(N-t)$ was "correct" for estimating lag-τ autocorrelations. The correct answer was that direct calculation of sample autocorrelations was a bad idea in any case and, given that the wrong estimate was being computed, the normalization was more-or-less irrelevant!

A. Multiple Window Estimates

Multiple window estimates of the spectrum [26] are a class of estimates based on approximately solving the integral equation that expresses the projection of $dX(f)$ onto the Fourier transform of the data $y(f)$. Taking the Fourier transform of the observed data, that is

$$y(f) = \sum_{t=0}^{N-1} x(t)e^{-j2\pi ft}$$

and using the spectral representation (1) for $x(t)$, we have the fundamental equation of spectrum estimation

$$y(f) = \int_{-1/2}^{1/2} K_N(f-\nu)\,dX(\nu) \tag{10}$$

where the Dirichlet kernel is given by

$$K_N(f) = \frac{\sin N\pi f}{\sin \pi f}e^{-2\pi f(N-1/2)}. \tag{11}$$

There are several points that must be remembered about this fundamental equation.

1) Because we may take the inverse Fourier transform of $y(f)$ and recover $x(t)$ for $0 \le t \le N-1, y(f)$ is a sufficient statistic and completely equivalent to the original data.
2) The finite Fourier transform $y(f)$ is not equivalent to the spectral generator $dX(\nu)$. Remember that $dX(\nu)$ is assumed to generate the entire data sequence for all t, not just the portion observed.
3) Despite definitions given in many elementary texts $(1/N)|y(f)|^2$, is *not* the spectrum, even in the limit of large N. It is the periodogram, biased and inconsistent.
4) While (10) is formally a convolution of dX with a Dirichlet kernel, it is more constructive to think of it as a Fredholm integral equation of the first kind. As such, it does not have a unique solution. It does, however, have useful approximate solutions. We mentioned above that "multiple window estimates" does not refer to a particular estimate, but rather to a class of estimates: the class is defined by the method used to form the necessarily approximate solution of the integral equation. Viewed in this way, spectrum estimation is in reality an inverse problem.

As multiple window methods have been described in [27] and in many papers, and since they are becoming the "standard" in geophysics [28], elaborate description of their properties is unnecessary; the reader is referred to [29]–[32]

for details, and we give only the equations necessary to define notation here.

B. Spectrum Estimation as an Inverse Problem

Recall the fundamental equation (10) and attempt eigensolution of the integral equation on the interval $(f-W, f+W)$ by assuming that the observable portion of dX has the expansion

$$d\hat{X}(f-\nu) = \sum_{k=0} x_k(t) V_k^*(\nu)\, d\nu \tag{12}$$

on the local frequency domain $(f-W, f+W)$. Here $V_k(\nu)$ is a Slepian function (discrete prolate spheroidal wave function, see [30, Appendix A] for definitions). Using the integral equation and properties of the Slepian functions we obtain the raw expansion, or eigencoefficients

$$y_k(f) = \sum_{t=0}^{N-1} e^{-j2\pi ft} \nu_t^{(k)}(N,W) x(t) \tag{13}$$

as the Fourier transform of the data $x(t)$ windowed by the kth Slepian sequence, $\nu_t^{(k)}(N,W)$. We retain the $K = 2NW$ coefficients corresponding to functions with eigenvalues $\lambda_k \approx 1$ for subsequent inference. These eigencoefficients represent the information in the signal projected onto the local frequency domain. This process resembles conventional, windowed spectrum estimation in that a fast Fourier transform may be used for efficient computation, but it differs in that standard estimates are best regarded as the first term of the multiple window expansion.

Because the Slepian sequences are time limited, they cannot be strictly bandlimited, and the kth sequence has a fraction $1-\lambda_k(N,W)$ outside the interval $(-W,W)$. Uncorrected, this out-of-band energy contributes bias which can be severe for the higher order, or transition eigencoefficients, that is those of order $k \approx K$. Among the various ways of dealing with this exterior bias, the best method found to date is by coherent sidelobe subtraction as outlined in [29]. Here we choose the bandwidth W to be large enough so that the bias on the lower order terms is negligible and $x_k(f) \approx y_k(f)$ for $k \ll K$, thereby estimating the higher order eigencoefficients by $x_k(f) \approx y_k(f) - \hat{b}_k(f)$ for larger k. The bias estimate $\hat{b}_k(f)$ is formed from an exterior convolution of the Slepian sequence with an estimate of (12) and iterated. Denote the estimated eigencoefficients by $x_k(f)$ and collect them in the vector $\boldsymbol{X}(f)$

$$\boldsymbol{X}(f) = [x_0(f), x_1(f), \ldots, x_{K-1}(f)]^T. \tag{14}$$

To see the dependence on the bandwidth W of the estimate, recall that there are $K \approx \lfloor 2NW \rfloor$ windows with eigenvalues near 1. If the spectrum is flat within the local domain, the coefficients are uncorrelated because the windows are orthogonal, and each contributes two degrees of freedom so estimates of the form $\boldsymbol{X}^\dagger(f)\boldsymbol{X}(f)$ have $2K$ degrees of freedom, where † denotes Hermitian transposition. If W is too small, we have poor statistical stability, but if W is too large, the estimate has poor frequency resolution. Typically W is chosen between $1.5/N$ and $20/N$ with a time-bandwidth product of four or five being a common starting point. Thus $W = 4/N$ or $5/N$ with corresponding $K = 6$ or 8 gives estimates with 12 or 16 degrees of freedom.

We must emphasize, however, that these only apply to the simplest forms of estimates and both quadratic inverse estimates (see [29]–[32]), and free parameter estimates of the type described therein give high-resolution estimates that are, within reason, largely independent of the choice of W. These estimates also give implicit extrapolations of the time series.

IV. High-Resolution Multiple-Window Spectrograms

Beginning with the estimate of $d\hat{X}$ defined in (12), define the narrowband process

$$X(t,f) = \int_{-W}^{W} e^{j2\pi t\xi}\, d\hat{X}(f \oplus \xi)$$

where $\oplus$ denotes addition with the constraint that $|\xi| < W$. On taking the inverse transform of the Slepian function, $X(t,f)$ becomes

$$X(t,f) = \sum_{k=0}^{K-1} \lambda_k \nu_t^{(k)} x_k(f)$$

where λ_k is the kth eigenvalue. Clearly the complex function $X(t,f)$ is not a time-frequency distribution but more akin to the output of a filter bank. Note first, however, that if we write the approximate impulse response of the implied filters, they go from maximum phase at $t = 0$ through zero phase at $t = (N-1)/2$, to minimum phase at $t = N-1$. Second, $X(t,f)$, as defined here extrapolates the signal to t outside the interval $[0, N-1]$ and so resembles the Papoulis estimates [33]. The squared amplitude $|X(t,f)|^2$ gives power as a function of time and frequency

$$F(t,f) = \frac{1}{K}\left|\sum_{k=0}^{K-1} \lambda_k x_k(f) \nu_t^{(k)}\right|^2 \tag{15}$$

and is an effective high-resolution spectrogram. Integrating this distribution over time gives the basic multiple window spectrum estimate

$$\sum_{t=0}^{N-1} F(t,f) = \frac{1}{K}\sum_{k=0}^{K-1} \lambda_k |x_k(F)|^2 \tag{16}$$

and so gives a much more accurate distribution of power than time-frequency distributions that simply match $|y(f)|^2$. Similarly, integrating $F(t,f)$ over frequency gives

$$\int_{-\infty}^{\infty} F(t,f)\, df = \frac{1}{K}\sum_{n=0}^{N-1}\left|\sum_{k=0}^{K-1} \lambda_k \nu_t^{(k)} \nu_n^{(k)}\right|^2 |x(n)|^2 \tag{17}$$

or, approximately, the convolution of $|x(t)|^2$ with $[\sin(2\pi Wt)/(\pi t)]^2$. Thus within a resolution interval $\Delta t = 1/(2W)$ power is approximately localized in time. Similarly, the narrowband properties of the Slepian sequences imply that cross terms are negligible for components separated in frequency by more than $2W$, so the time-frequency resolution area is of order one. There are, however, two problems with this estimate: first, its distribution is proportional to χ_2^2, so it is statistically unstable; second, in common with time-frequency distributions of the form $x(t) \cdot \overline{y}(f)$ (see [20, ch. 14]), this estimate can be thought of as

$$\int \gamma_L(\xi, f) e^{j2\pi\xi t}\, d\xi$$

and so has "mixed" continuity properties caused by integrating across the expected δ-function sheets at 45 degrees in only one of the two variables. A more serious criticism is that, although such estimates satisfy enhanced marginal conditions, they appear to overlook the essential feature of correlation between frequencies more than $2W$ apart. Nonetheless, this "high-resolution" spectrogram represents, in applications where the spectrogram is useful, a vast improvement on the standard version. This estimate, like other multiple window estimates, can obviously be extended to include overlapping data sections, so high-resolution spectrograms of long data sets can be formed by averaging the above estimates. Much more, indeed, is possible. Extending the definition (13) to make the base position b explicit

$$y_k(b, f) = \sum_{n=0}^{N-1} e^{-j2\tau fn} \nu_n^k(N, W) x(b + n) \tag{18}$$

we have, corresponding to (15)

$$F(b \oplus t, f) = \frac{1}{K} \left| \sum_{k=0}^{K-1} x_k(b, f) \nu_t^{(k)} \right|^2 \tag{19}$$

where $\oplus$ again represents a restricted sum, with $0 \leq t \leq N - 1$. (Given the extrapolation properties of these estimates, mentioned above, this restriction is not strictly necessary, only conservative.) Section IV-A, on nonstationary quadratic-inverse estimates, discusses another way to improve on the standard spectrogram, and the Section IV-B shows how the basic expansion $X(t, f)$ may be used to estimate correlations between frequencies.

A. Nonstationary Quadratic-Inverse Theory

The problem of stability in the above estimate can be "solved" by quadratic-inverse theory [30]–[32]. This is a way to generate minimum-variance unbiased estimates of second-moment quantities directly from the eigencoefficients of the linear inverse solution without going through the *ad-hoc* procedure of generating the linear inverse, squaring, and then estimating the required second moments from these. Here we compute the eigensequences of the squared kernel (rigorously, the squared truncated kernel)

$$\alpha_l A_l(n) = \sum_{m=0}^{N-1} \left[\frac{\sin 2\pi W(n - m)}{\pi(n - m)} \right]^2 A_l(m). \tag{20}$$

These sequences rapidly approach those of the continuous time problem [34], and there are approximately $4NW$ nonzero eigenvalues. Thus we have approximately $\alpha_l \sim 2NW - l/2$ for $l = 0, 1, \ldots, 4NW$ and, as the variances of the quadratic-inverse coefficients are proportional to α_l^{-1}, the first few coefficients are nearly as stable as the standard multiple window spectrum. The associated bases matrices

$$A_{jk}^{(l)} = \sqrt{\lambda_j \lambda_k} \sum_{n=0}^{N-1} \nu_n^{(j)} \nu_n^{(k)} A_l(n) \tag{21}$$

are real, symmetric, and trace-orthogonal; that is

$$\mathrm{tr}\left\{ \boldsymbol{A}^{(l)} \boldsymbol{A}^{(m)} \right\} = \alpha_l \delta_{lm}. \tag{22}$$

The expansion coefficients corresponding to $F(t, f)$ are

$$\hat{P}_l(f) = \frac{1}{\alpha_l} \boldsymbol{X}^\dagger(f) \boldsymbol{A}^{(l)} \boldsymbol{X}(f) \tag{23}$$

and so we have

$$P(t, f) = \sum_{l=0} \hat{p}_l(f) A_l(t). \tag{24}$$

The coefficients $\hat{p}(f)$ are often informative in their own right (see [31] and [35]). In particular, the zero-order function $A_0(t)$ is approximately constant so $\boldsymbol{A}^{(0)} \approx \boldsymbol{I}$, and $\hat{p}_0(f)$ is approximately standard multiple-window spectrum. The order-one function $A_1(t)$ is approximately equal to $t - (N - 1)/2$. Thus $\boldsymbol{A}^{(1)}$ is zero on the diagonal and approximately constant on the sub- and superdiagonal, and $\hat{p}_1(f)$ is approximately the first time-derivative of the spectrum, and so on. One useful *ad-hoc* quantity is $\hat{p}_1(f)/\hat{p}_0(f)$, approximately the time-derivative of $\ln S(t, f)$. For example, in [35] $\hat{p}_1(f)$, computed from residuals of a global temperature series from 1854–1992, was almost uniformly negative across frequencies. While one must consider the series formally as nonstationary, the most reasonable explanation is not metaphysical, but simply that instrumentation and spatial coverage has improved since 1854. In this example the quadratic inverse estimates are preferable to a spectrogram or the Loève spectrum; the decrease in power is relatively small and the data series has only 138 samples, so computing a spectrogram would be difficult and could be easily misinterpreted. Here, the negative derivative of the noise spectrum probably reflects little more than the improvements in instrumentation and spatial coverage that have occurred since 1854.

Expanding $F(t, f)$ of (16) in terms of the $A_l(t)$'s, it can be seen that the resulting coefficients $F_l(f)$ are biased by α_l/K. However $F(t, f)$ is biased and positive, whereas truncation and Gibb's ripples can cause $P(t, f)$ to be negative.

Although spectrograms are insensitive to correlations between widely different frequencies, when the temporal evolution of the spectrum is slow, spectrograms form a useful intermediate class of time-frequency distributions. Quadratic-inverse estimates improve on the spectrogram by allowing for changing power within the block, and for tests between blocks.

While the basic theory of multiple-window methods is usually written in terms of a finite block size N, we can obviously apply the same methods to overlapping time blocks to form spectrograms [29]. On each block we estimate a dynamic spectrum $D(t,f)$, its frequency derivative $D'(t,f)$ (from [30]), time derivative $\dot{D}(t,f)$ (from [23]), and perhaps higher terms. These low-order terms are very stable with variances proportional to $1/\alpha_l$. Because $\hat{p}_0(f)$ is approximately the spectrum, $\hat{p}_1(f)$ is approximately the first time-derivative of the spectrum, etc. We can either make a "smoother" that uses these or, better, given $D(t,f)$ and $\dot{D}(t,f)$, we can test if an estimate $D(t+\Delta,f)$ is "reasonable." Also, the "Nyquist sampling rate" for $F(t,f)$ is simply $\Delta = 1/(2W)$, so K samples spaced N/K are obtained in each time block. Thus, if the blocks are offset by Δ we have K estimates at each point of the time-frequency plane, so that averages and variances can be computed. The covariances between blocks can be computed, tests for homogeneity of correlated variances are known, so the procedure can be used to test whether a choice of N and W is reasonable. Assuming that the estimate is reasonable, note that the average of the $F(t,f)$'s at each resampling time will be reasonably stable. Because of correlations between blocks, the stability of an average will be much less than $2K$ degrees-of-freedom, but the long lower tails characteristic of $\log\chi_2^2$ distributions are considerably suppressed. We use log spectra because: formally, the information content of a signal is measured by its Wiener entropy- a logarithmic measure; pragmatically, most engineering applications are designed for human use and both the eye and ear have a logarithmic response. With the exception of helioseismology, it is difficult to find plots of power spectra which are not on a logarithmic (or decibel) scale. Thus we have a spectrogram with both good stability and time resolution.

Incidentally, taking either a log spectrogram or $(\partial/\partial t)\ln D(t,f)$ into a singular value decomposition and then analyzing the time eigenvectors as standard time series is often very useful (see [29]).

B. Multiple Window Estimates of the Loève Spectrum

Taking the complex demodulates at two different frequencies f_1 and f_2, an obvious estimate of their covariance is

$$\hat{\gamma}(f_1,f_2) = \frac{1}{K}\sum_{t=0}^{N-1} X(t,t_1)X^*(t,f_2) \qquad (25)$$

where the normalization is proportional to the number of independent samples. The orthogonality of the Slepian sequences gives

$$\hat{\gamma}(f_1,f_2) = \frac{1}{K}\sum_{k=0}^{K-1} x_k(f_1)x_k^*(f_2). \qquad (26)$$

This is the estimate given in [26] and generally works well (see [36], [37] or [38]). An alternative motivation is that, if we consider the product of two estimates of the form

$$dX(f\oplus\xi) \sim \sum_{k=0}^{K-1} \hat{x}_k(f)V_k(\xi)\,d\xi \qquad (27)$$

for $|\xi| < W$ then, guided by the continuity arguments of Section III, we use a weight $W(\xi_1,\xi_2) = \delta(\xi_1-\xi_2)$ so that smoothing over a bandwidth W is done on the stationary frequency, and no smoothing on the nonstationary direction, the same estimate is obtained. A similar smoothing scheme was proposed in [39] and applied effectively in [40].

It is often useful to plot the dual-frequency spectrum as a dual-frequency coherence, that is, defining

$$C(f_1,f_2) = \frac{\hat{\gamma}(f_1,f_2)}{[S(f_1)S(f_2)]^{1/2}} \qquad (28)$$

plot a dual-frequency magnitude-squared coherence $|C(f_1,f_2)|^2$ and the phase. Significance level calculations for this magnitude-squared coherence (MSC) are exactly the same as they are for ordinary MSC calculations (see [41]).

There are far too many extensions of this approach to describe in detail here; however, an indication of some directions should be mentioned.

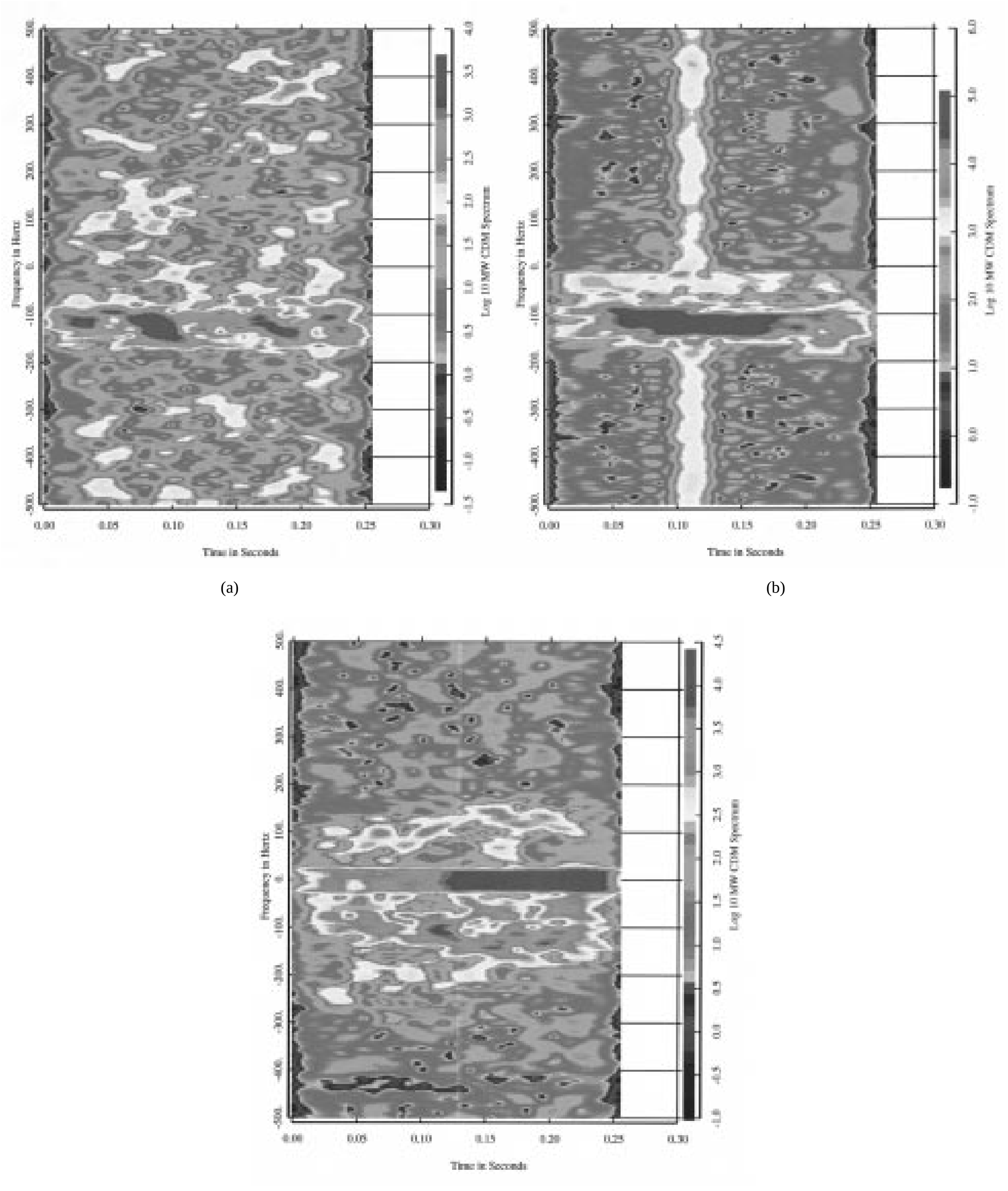
1) The correlation estimate in (26) can be extended to include a time delay, that is ave$\{X(t,f_1)(X^*(t+\tau,f_2))\}$ which results in a quadratic form

$$\hat{\gamma}(f_1,f_2,\tau) = \sum_{j=0}^{N-1}\sum_{k=0}^{N-1} x_j(f_1)x_k^*(f_2)B_{jk}(\tau). \qquad (29)$$

2) We may use a similar quadratic form with, for example, $\boldsymbol{A}^{(1)}$, to test for energy transfer between frequencies. More generally, for a specific spectral pattern of interest, a weight $W(\xi_1,\xi_2)$ is chosen to emphasize it, and the integration over $-W < \xi_1,\xi_2 < W$ results in an appropriate weight matrix.
3) Treat $X(t,f)$ as a matrix, possibly scaling by $1/S(f)^{1/2}$, compute its singular value decomposition (SVD), then treat the dominant time-eigenvectors as new time series.
4) The same procedures can be applied to multivariate time series; in bivariate problems compute ave$\{X(t,f_1)Y^*(t,f_2)\}$ or similar, or several series can be "stacked" in the SVD process.
5) In communications signals, it is common to encounter the same signal with sidebands reversed. In this case the appropriate smoother would be perpendicular to the standard one and, as in [26], can be obtained by leaving the second coefficient unconjugated.

V. Spectrum Analysis of Radar Signals

We now apply these ideas to three radar data sets: sea clutter (i.e., radar backscatter from an ocean surface) on its own, weak target signal in clutter, and strong target signal in sea clutter. The target signal was due to the echo



(a)

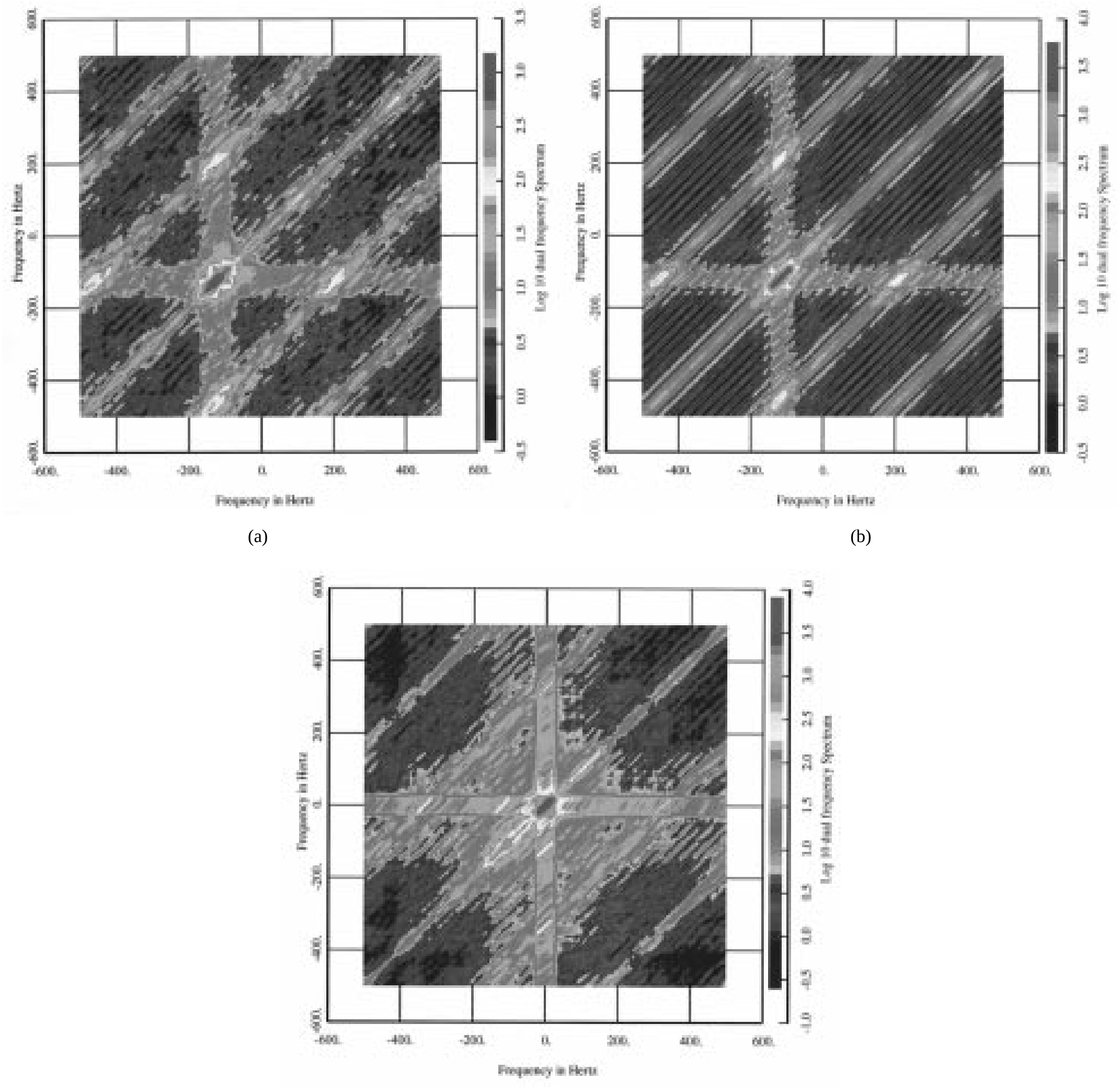
(b)

(c)

Fig. 1. Spectograms. (a) Radar clutter only data set, HH channel, eight windows, three sections, step ten. (b) Radar weak growler data, HH channel, eight windows, three sections, step ten. (c) Radar strong growler data, HH channel, eight windows, three sections, step ten.

from a small piece of ice "growler" floating in the ocean under the dynamics of the ocean waves. Note that these are actual data, not simulations, and consequently the clutter components in all three series are necessarily different and the energy in the clutter component of the series varies with conditions. Each series consists of 256 complex samples taken at $\Delta T = 1.0$ ms.

Fig. 1(a)–(c) shows high-resolution spectrograms of the data computed by the method of Section IV but averaging the results of ten sections of 229 samples each, each offset



(a)

(b)

(c)

Fig. 2. Loève transform. (a) Radar clutter only data, set, HH channel, eight windows, three sections, step ten. (b) Radar weak growler data, HH channel, eight windows, three sections, step ten. (c) Radar strong growler data, HH channel, eight windows, three sections, step ten.

by three samples. A time-bandwidth product of six was used with $K = 10$ windows on each section. The bandwidth is thus $\pm 6/(229\Delta T) = \pm 26$ Hz. The section offset used here is smaller than recommended above and the section averaging used to suppress the lower tails of the $\log\, \chi^2$ distribution. (Normally, we would not attempt to compute a spectrogram from a sample of size 256.) Fig. 1(a), the spectrogram of the clutter, shows a band near -110 Hz with more power than elsewhere, but otherwise the spectrogram is reasonably flat over the clutter spectrum. The contribution due to receiver noise is about 20 dB below the clutter spectrum.

By contrast, Fig. 1(b), the spectrogram of the weak growler, shows a strong, frequency-independent vertical stripe (due to clipping of the time series) near $t = 27$ ms as well as a second frequency stripe at about -25 Hz in addition to the features visible in the clutter spectrogram.

With the strong target, Fig. 1(c), the stripe centered near 0 Hz (representing the Doppler shift of the target signal) is much more obvious, the clutter band is still there, and there is a weaker clutter image band, possibly due to a slight imbalance in the in-phase and quadrature channels of the coherent receiver.

Fig. 2(a)–(c) shows the estimates of the corresponding Loève spectra, and the gain in information beyond that apparent in the spectrogram is striking. First, the diagonal bands evident in all three series show that the data are periodically correlated, or cyclostationary; this was not obvious in the spectrograms, nor expected. In Fig. 2(a) it can be seen that the peak of the Loève spectrum for clutter is near -125 Hz, as before. In Fig. 2(b), the weak target signal case, an extra peak centered at about -25 Hz, while in Fig. 2(c), the strong target signal case, the peak near zero is dominant but, in contrast to the spectrogram, the periodic correlation of the clutter is still visible.

Fig. 3(a)–(c) presents the corresponding WVD of the sea clutter on its own, weak target in sea clutter and strong target in sea clutter, respectively. The important feature to observe here is the presence of a zebra-like pattern (alternating between dark and bright narrow stripes) in the images due to the presence of a target signal. This pattern occupies an area located between the instantaneous frequency plot of the target near 0 Hz and that of the clutter. This pattern is indeed a manifestation of the cross WVD terms due to the combined presence of a target signal and clutter. Most importantly, the presence of this zebra-like pattern is found to be 1) fairly pronounced at relatively low target signal-to-clutter ratios and 2) relatively robust to variations in the target signal-to-clutter ratio [17].

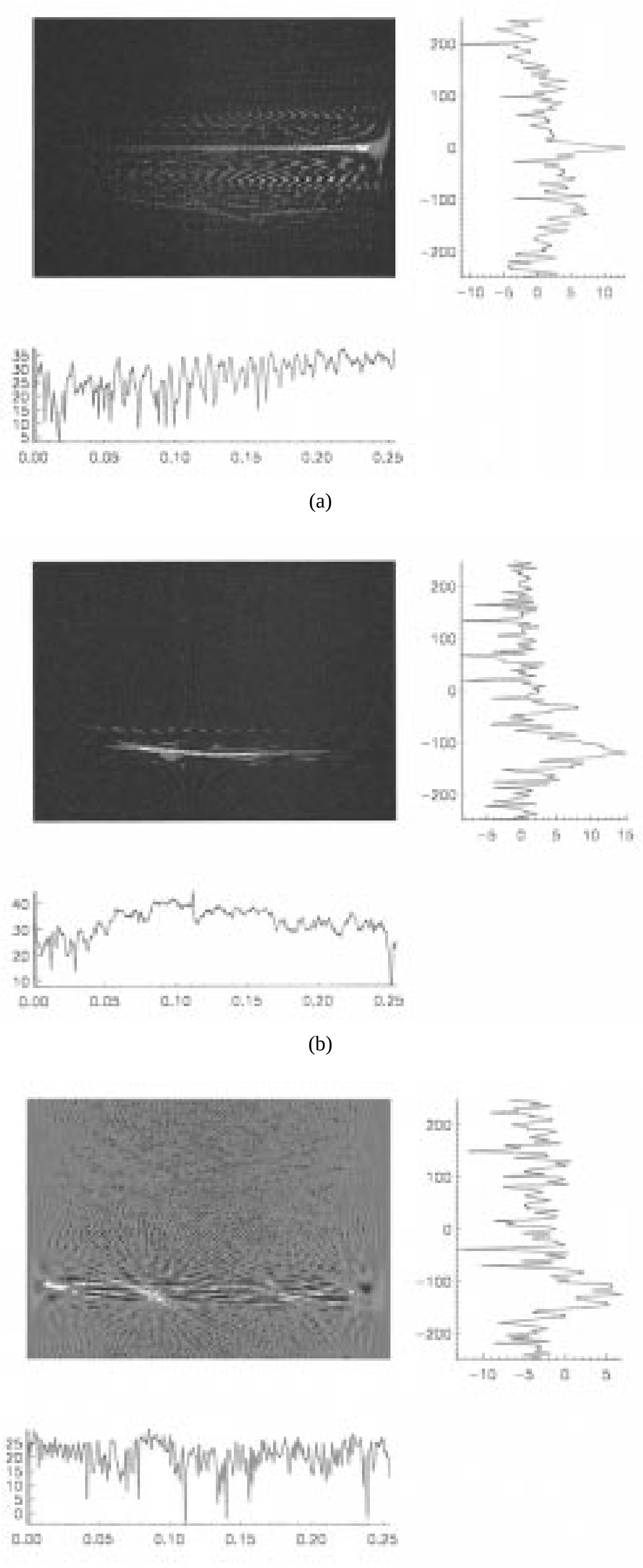
Although the high-resolution spectra estimates of Figs. 1 and 2 based on the method of multiple windows display the dynamic spectrum of the radar signals in ways that are similar (in some parts) and yet different (in other parts) from the corresponding WVD of Fig. 3, the important point to note from these two differently computed sets of images is that both approaches accentuate the differences between the different classes of radar signals in their own individual ways, making them more visible than the original time series. Simply put, the power of these methods lies in their ability to make a weak target signal buried in a strong clutter background visible in signal-processing terms.

VI. Modular Learning Machine for Adaptive Signal Detection

Regardless of whether we use the high-resolution images exemplified by Figs. 1 and 2 or the WVD of Fig. 3, these two approaches do share a common property: the one-dimensional time series representing the received signal is transformed into a highly redundant 2-D image. For the detection strategy to be computationally efficient, the redundant information contained in this image would have to be removed by some means. This is not so different from signals such as speech where redundancy is stripped for coding and then added later for error protection.

In pattern recognition theory, the removal of redundant information is referred to as feature extraction [42], [43]. Traditionally, feature extraction is followed by pattern classification. At the output of the pattern classifier a decision is made as to whether the image applied to the feature extractor, or equivalently the original received signal, belongs to one of two possible (hypotheses) classes.

1) *Null Hypothesis*, H_0: The received signal consists of noise alone.



(c)

Fig. 3. (a) WVD for a clearly visible growler. (b) WVD for a barely visible growler. (c) WVD for sea clutter. For the images, horizontal axes: time in seconds; vertical axes: frequency in Hz. Horizontal axes of power spectra: in dB.

2) *Other Hypothesis*, H_1: The received signal consists of a target signal plus noise.

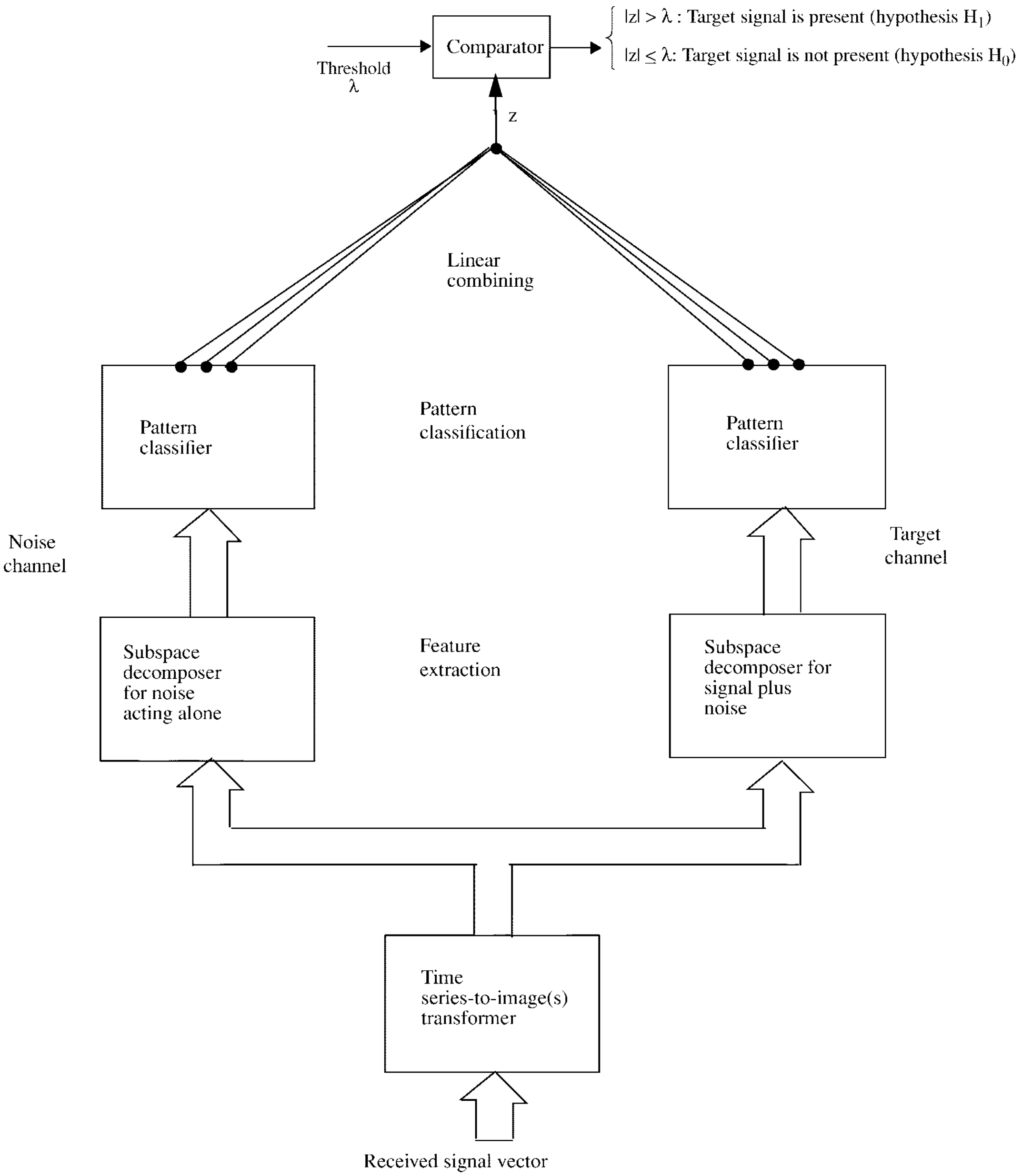

Comparator
|z| > λ : Target signal is present (hypothesis H_1)
|z| ≤ λ: Target signal is not present (hypothesis H_0)
Threshold λ
z
Linear combining
Pattern classifier
Pattern classification
Pattern classifier
Noise channel
Target channel
Subspace decomposer for noise acting alone
Feature extraction
Subspace decomposer for signal plus noise
Time series-to-image(s) transformer
Received signal vector

Fig. 4. Block diagram of the two-channel receiver.

In other words, we have a binary hypothesis testing problem, the solution of which is optimized in some statistical sense.

Fig. 4 shows the block diagram of a *modular learning machine* [17], [44] for adaptive signal detection, based on the strategy described. At the input end of the machine we have a time series-to-image transformer that uses dynamic spectrum analysis or time-frequency analysis for its design. From that point on, the machine splits into two channels, one termed the noise channel and the other termed the target channel. The two channels are linearly combined at their outputs, and then a final decision is made as to whether a target signal is present in the received signal or not.

Each channel consists of two functional blocks: feature extractor and pattern classifier. A popular method for implementing the feature extractor is principal components analysis (PCA), the aim of which is to learn the dominant eigenvectors that are most representative of the different realizations of the pertinent class of data. In basic mathematical terms, this operation involves performing an eigendecomposition on the covariance matrix of a data vector obtained by scanning the image on a column by column basis, arranging the eigenvalues in decreasing order, and retaining only those eigenvectors that are associated with the dominant eigenvalues. In effect, the PCA performs a subspace decomposition on the images belonging to class H_0 or class H_1 and converges onto a solution that captures the features that are most common to the different physical realizations of the class in question. Accordingly, when an image representing class H_0, for example, is projected onto the two subspaces computed by the noise and target channels, the outputs of the feature extractor in the noise channel are relatively small, whereas the corresponding outputs of the feature extractor in the target channel are relatively large. The reverse holds for the presentation of an image representing class H_1. We may therefore state that the feature extractor in the noise channel is adaptively matched to input data known to consist of noise alone. A similar statement holds for the feature extractor in the target channel. Both feature extractors are designed using a self-organized learning procedure that tracks the statistical variations in the received signal.

Turning next to the pattern classifiers and linear combiner, they are designed by using a supervised learning procedure. To do this, we may use one of two procedures.

1) The two pattern classifiers are trained separately, with hard decisions being made at their respective outputs. When the machine is presented with a received signal known to consist of noise only, the pattern classifier in the noise channel is trained to classify that signal, for example, as belonging to one of the following three categories:

   a) the received signal is definitely noise;
   b) the received signal is on the border line of being interpreted as noise or a weak target signal plus noise;
   c) The received signal looks like a weak target signal plus noise.

   Correspondingly, when the machine is presented with a received signal known to consist of target signal plus noise, the pattern classifier in the target channel is trained to classify that signal as belonging to one of the following three categories:

   a) the received signal contains a strong target signal;
   b) the received signal contains a weak target signal;
   c) the received signal looks like noise.

   The outputs of the two channels are linearly combined to produce the overall output z, where the final decision (in favor of hypothesis H_0 or H_1) is made. To design the linear combiner, the machine goes through another round of supervised training with data not seen before. The categorization of the noise/target channel output in the manner described here may be viewed as a discrete approximation to soft-decision making.

2) The two pattern classifiers and linear combiner are all trained simultaneously. The individual outputs of the two pattern classifiers (three each, for example) are now free to assume values set by the activation functions of their output neurons (processing units).

The attractive feature of method 1) is that the decision boundaries between the different subclasses of the received signal are well defined at the outputs of the two channels. However this advantage, if any, over method 2) is gained at the expense of increased training.

Irrespective of whichever of these two methods is adapted, the adaptive two-channel receiver of Fig. 4 relies on the use of learning-to-learn procedures for its design. Most importantly, the target and noise channels tend to reinforce each other in their individual decisions, thereby providing for an improved detection performance over that attainable with a single channel. To that end, the use of linearly combining the outputs of the two channels, viewed as a form of ensemble averaging, is usually considered to be superior to the use of majority voting [45].

A. How Does the Adaptive Receiver of Fig. 4 Respond to a Nonstationary Environment?

An issue that may need further clarification is how the adaptive receiver of Fig. 4 is able to respond to statistical variations in the environment. The answer to this fundamental issue lies in: 1) the transformation of a time series into an image or images; 2) the adaptive subspace decompositions of the images so computed; and 3) the training of the pattern classifiers under the tutelage of a "teacher."

The time series-to-image transformers enhance the nonstationary character of the received signal, making it more visually discernible and therefore more readily learnable. In this context, it is noteworthy to recognize that the sonar-based echolocation system of a bat relies on the computation of time-frequency maps for its own operation [46], [47]. Indeed, a bat is able to detect and track its prey (e.g., a flying insect) in a difficult environment with a facility and success rate that would be the envy of a sonar and radar engineer.

The adaptive subspace decompositions performed on the time-frequency images (maps) provide for compact representations of a wide variety of different realizations of the two classes of data, that is, under hypotheses H_0 and H_1. This is done by transforming the higher dimensional spaces of the images to lower dimensional spaces, subject to the constraint of information preservation (i.e., reconstruction of the original data with minimal distortion). Moreover, the features constituting the compact representations are decorrelated, thereby paving the way for the efficient training of the pattern classifiers.

Finally, it is in the design of the pattern classifiers where the accounting for nonstationary behavior of the environment comes into focus. Ground-truthed data pertaining to hypotheses H_0 and H_1 are presented alternately to the receiver, and the free parameters of the pattern classifiers are adjusted so as to minimize a statistical criterion of interest. By "ground truthing" we mean that when the data are collected, the prevalent environmental conditions (i.e., the target is present or not) are carefully monitored and recorded by human observers. In the course of this supervised training, information contained in the input data is transferred and stored in the values assigned to the free parameters of the subspace decomposers and pattern classifiers. The net result of the training process is that a nonlinear decision boundary is constructed in the input space between the hypotheses H_0 and H_1, with the data having a direct say in how the decision boundary is adaptively constructed. On the one side of the decision boundary we have data points assigned to hypothesis H_0; each such point represents a different realization of this hypothesis knowing that no target is present, with the difference arising because of statistical variations in the environment. On the other side of the boundary we have data points assigned to the hypothesis H_1; each such data point represents a different realization of this second hypothesis knowing that a target is present, with the difference again being due to statistical variations in the environment. Subject to the proviso that the training data are representative of the nonstationary environment, and the pattern classifiers are therefore forced to assign the data points appropriately to class H_0 or class H_1, under the tutelage of a teacher, the receiver should achieve a performance under test that is close to the performance under training. It is assumed here that the test data are different from the training data but drawn from the same environment. Indeed, we can make the following general observation [45]: the more exhaustive the training data set is in its representation of the nonstationary environment, the more likely it is that the receiver adapts to its environment fully and therefore be able to exhibit a robust detection performance.

The receiver design described herein differs markedly from the way in which classical receivers are designed. In the classical approach, we start with a mathematical model of the received signal and end up testing the receiver performance with real-life data. The success of the classical approach rests largely on how close the mathematical model is to the realities of the data. On the other hand, in the adaptive approach through learning described in this paper, the data set is allowed to "speak for itself" in how the receiver is designed. Stated in yet another way, the classical approach relies on mathematical tractability, and the challenge in this approach is to formulate the right mathematical model for the received signal that accounts for the nonstationary behavior of the environment. In the adaptive approach through learning, the need for mathematical modeling is eliminated, and the challenge in this modern approach is twofold: 1) collect a ground-truthed database that is large enough in size and fully representative of the nonstationary environment and 2) design the receiver using appropriate learning-to-learn procedures that respond to statistical variations of the environment.

VII. Case Study: Radar Target Detection of a Small Target in Sea Clutter

The adaptive two-channel receiver of Fig. 4 defies a statistical analysis of detection performance along traditional lines due to its nonlinear nature. Therefore, to evaluate the practical merit of this new receiver, we performed a case study involving the detection of a growler floating in an ocean environment. A growler is a small piece of ice that is broken off an iceberg. The above-surface visible portion of it is about the size of a grand piano (i.e., a radar cross-section of about 1 m^2). However, recognizing that about 90% of the volume of ice lies below the water surface, a growler represents an object large enough to be hazardous to navigation in ice-infested waters, such as those encountered on the East Coast of Canada during the Spring and early Summer. The radar task at hand is that of detecting the radar echo from a growler in the presence of interference represented by sea clutter.

For the collection of radar data representative of this environment, an instrument-quality radar system called the IPIX radar was used. The IPIX radar [48] is a fully coherent, polarimetric, X-band radar system equipped with computer control and digital data acquisition capability. The present study is confined to the use of coherent data collected under the polarimetric condition of horizontal transmit and horizontal receive only. The radar was operating in a staring mode (i.e., pointing onto a patch of the ocean surface). A series of experiments using the IPIX radar was performed at a site located on the East Coast of Canada. The radar was mounted at a height above sea level that would be representative of a ship-mounted radar. Ground-truthing of the data collected was maintained throughout the experiments, thereby providing knowledge of the conditions under which the various datasets were collected. This case study was chosen for the application at hand because both the target of interest (a growler) and the background interference (sea clutter) are known to exhibit nonstationarity, which would require the use of adaptivity. Moreover, the generation of sea clutter is governed by a nonlinear dynamical process, which would therefore require the use of nonlinear processing. Thus, the detection of a growler in sea clutter provides a suitable medium for testing the capabilities of our new detection strategy. Details of this case study were presented in [17]. The material presented here is a summary of the results presented in that paper.

A. Details of the Receiver

The 2-D WVD image used in the study had a time dimension $L = 256$ with spacing $T = 1$ ms, and a frequency dimension $M = 256$ with spacing $F = 4$ Hz. Note that the WVD image is real valued even though the received signal is complex valued.

Each of the two PCA networks in Fig. 4 consisted of a feedforward NN with an input layer of $M = 256$ source nodes (fed from the WVD image of the received signal) and a single computation layer of $p = 5$ linear neurons. Both networks were fully connected, in that each neuron of either network was connected to all the source nodes of its respective input layer. The total number of connections/independent weights for each PCA network was 1280. The training set for PCA$^{(0)}$ network was made up of $A_0 = 2000$ epochs, representing hypothesis H_0. The training set for PCA$^{(1)}$ network was made up of $A_1 = 50$ epochs, representing hypothesis H_1. The individual epochs of WVD images were generated using examples of the received signal, each being made up of 256 samples. Both PCA networks were trained using the generalized Hebbian algorithm (GHA) due to Sanger [49]. Let $\boldsymbol{x}(t)$ denote the input vector applied to the algorithm at iteration (time step) $t, \boldsymbol{y}(t)$ denote the corresponding value of output vector, and

$\boldsymbol{W}(t)$ denote the weight matrix of the PCA network under training. The change $\Delta\boldsymbol{W}(t)$ in the weight matrix computed by the algorithm at iteration t is defined by

$$\Delta\boldsymbol{W}(t) = \eta\big(\boldsymbol{y}(t)\boldsymbol{x}^T(t) - LT[\boldsymbol{y}(t)\boldsymbol{y}^T(t)]\boldsymbol{W}(t)\big) \quad (30)$$

where the superscript T denotes matrix transposition, the operator LT makes the matrix enclosed inside the square brackets lower triangular by setting all the elements above its diagonal equal to zero, and the scaling factor η denotes the learning rate parameter of the algorithm. The GHA operates on the WVD directly. Most importantly, it is well suited for the application at hand by virtue of the large size of the training data and the ability of the algorithm to track changes in the input data from one epoch to the next.

Turning next to the pattern classifiers in Fig. 4, we used a type of feedforward NN's known as multilayer perceptrons (MLP's). The input layer of each MLP consisted of an array of $p \times L$ source nodes fed from a compressed image with $p = 5$ and $L = 256$. The first hidden layer consisted of an array of 5×15 neurons. The network architecture of this layer was constrained as shown in Fig. 5. Specifically, it incorporated the following concepts for improved training and perhaps better generalization performance [13], [45].

1) *Receptive Field*: This means that a neuron in each row of the first hidden layer is connected only to a certain number of source nodes (denoted by R) that lie in its local neighborhood in the corresponding row of the input layer.
2) *Overlap of Receptive Fields*: This means that the receptive fields of adjacent neurons in a particular row overlap by a certain number of source nodes (denoted by S).
3) *Weight Sharing*: This means that the receptive fields of all the neurons in a particular row share the same set of synaptic weights.

For our present study we chose $R = 32$ and $S = 16$. In addition, each MLP had a second hidden layer of 25 neurons and output layer of three neurons, both of which were fully connected. The neurons in both MLP's were all nonlinear, using a sigmoid activation function defined by the logistic function

$$\varphi(\nu) = \frac{1}{1 + \exp(-\nu)} \quad (31)$$

where ν is the induced local field of the neuron. The induced local field ν includes a threshold (negative of bias), which is represented by an adjustable weight connected to an input fixed at −1. A total of 10 156 examples of the received signal were used to do the supervised training of the MLP's. They were made up as follows: $B_0 = 7150$ examples representing hypothesis H_0, and $B_1 = 3006$ examples representing hypothesis H_1. Each example of the received signal was 256 samples long. This training dataset was completely different from that used to train the PCA networks. The two MLP's and linear combiner were treated as a single entity and trained using the back-propagation (BP) algorithm [45], [50]. The BP algorithm operates in two phases. In Phase I, called the forward phase, the synaptic weights of the network are fixed. In this phase the signal applied to the input layer propagates through the network in a forward manner, layer by layer. Phase I is completed by calculating the error signals, defined as the difference between the elements of the desired response vector and the corresponding values of the actual output signals of the neurons in the output layer. The error signals are propagated through the network in the backward direction in Phase II, called the backward phase. In particular, they are used in a generalized delta rule (i.e., generalization of the popular least mean square (LMS) algorithm) to compute the adjustments applied to the individual synaptic weights of the multilayer perceptron. The BP algorithm has two useful properties:

1) simplicity of implementation;
2) stochastic gradients (i.e., derivatives of the error performance surface with respect to the weights in the network).

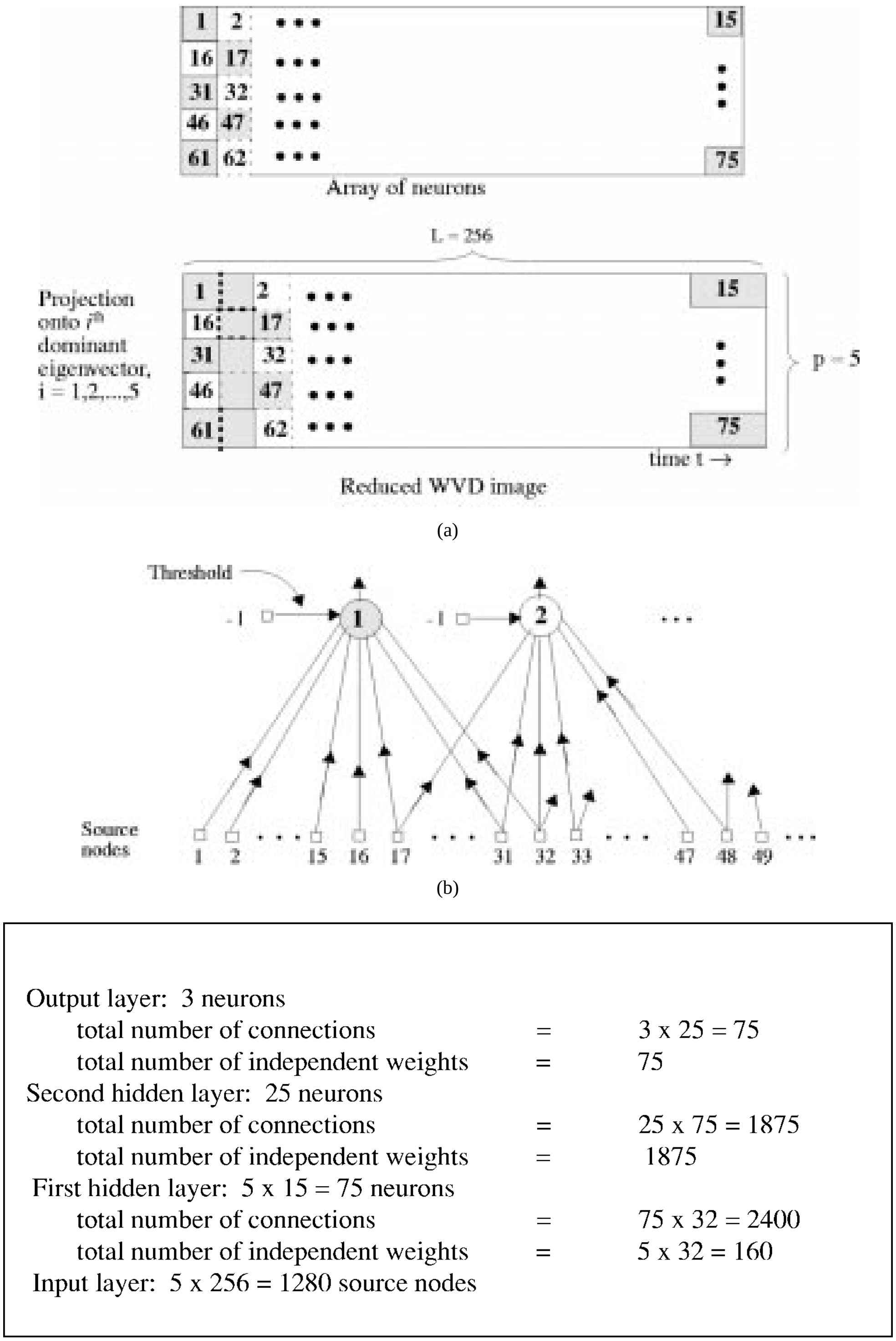
It is because of these two properties that the BP algorithm has established itself as the workhorse for the training of MLP's intended particularly for pattern classification, hence its use for the design of the pattern classifiers in the two-channel receiver of Fig. 4.

B. Detection Results

Fig. 6 presents a visual display of the predetection performances of two different receivers.

1) Doppler constant false-alarm rate (CFAR) receiver; this system was chosen as a frame of reference because of its widespread use as a conventional radar receiver.
2) NN implementation of the two channel receiver of Fig. 4, which involved the use of three output nodes per channel.

Predetection refers to the receiver output prior to thresholding. The results of Fig. 6 were obtained for a long dwell time (approximately 35 s) along a range swath of 200 m and a range gate (resolution) of 5 m; the total number of radar samples represented here is 2.68×10^6. The test data used here were completely different from the data used to train the PCA networks and those used to train the MLP's. The darkness of the display in Fig. 6 is a measure of the actual power of the receiver output before thresholding. The two parts of the figure have been normalized separately to remove any bias introduced by changes in dynamic ranges of the receivers. The Doppler CFAR and NN receivers paint the two hypotheses H_0 and H_1 in dramatically different colors. In particular, the discrimination between the clutter background (hypothesis H_0) and target (hypothesis H_1) is far more pronounced in the NN receiver than it is in the Doppler CFAR receiver. This is a direct result of the fact that the Doppler CFAR receiver is basically linear, whereas the NN receiver is highly nonlinear thereby capturing the information content of the received signal to a fuller extent.



(a)

(b)

Output layer: 3 neurons
total number of connections = 3 x 25 = 75
total number of independent weights = 75
Second hidden layer: 25 neurons
total number of connections = 25 x 75 = 1875
total number of independent weights = 1875
First hidden layer: 5 x 15 = 75 neurons
total number of connections = 75 x 32 = 2400
total number of independent weights = 5 x 32 = 160
Input layer: 5 x 256 = 1280 source nodes

(c)

Fig. 5. Architectural details of each MLP. (a) 2-D display of first hidden layer; threshold is the negative of bias. (b) First row of neurons in the first hidden layer. (c) Connectivity of each MLP.

To further emphasize the performance difference between these two receivers, Fig. 7 shows their postdetection performances obtained by comparing the amplitude of each receiver output against a threshold. The threshold was set for a false alarm rate of 10^{-3}; that is, the probability that a target is present in the received signal when actually it is not was prescribed not to exceed 10^{-3}. This false alarm rate is considered typical for the operation of a surveillance radar. The color black in Fig. 7 signifies the presence of the growler (hypothesis H_1), and white signifies its absence (hypothesis H_0). With the radar operating in a dwelling mode, the growler should ideally be visible to the radar all of the time, that is, we should ideally see a continuous black strip extending all along the time axis. With this ideal

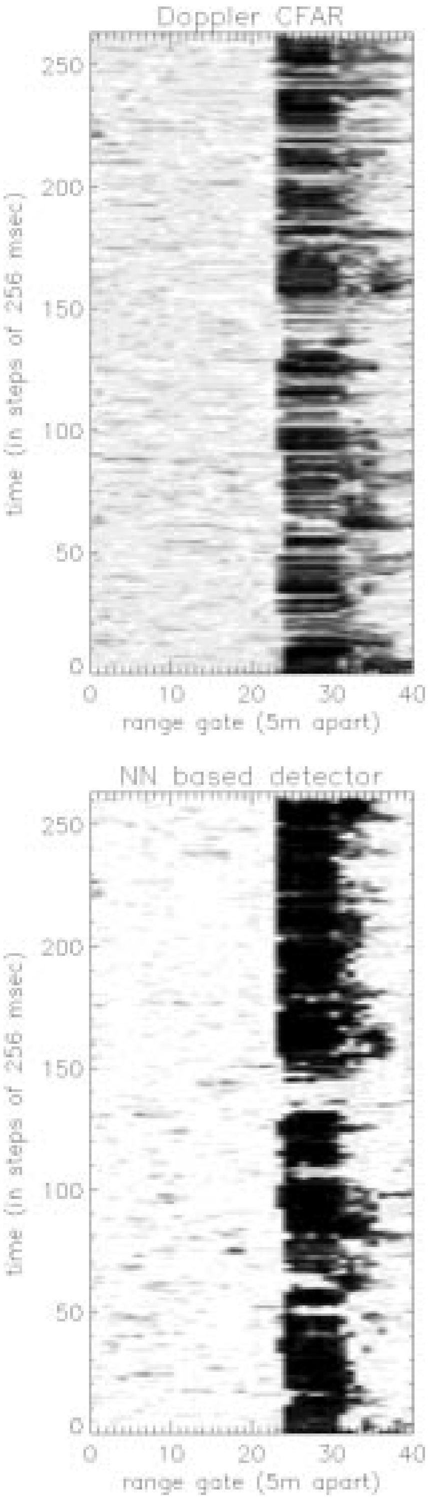


Fig. 6. Predetection statistics for the Doppler CFAR and NN receivers.

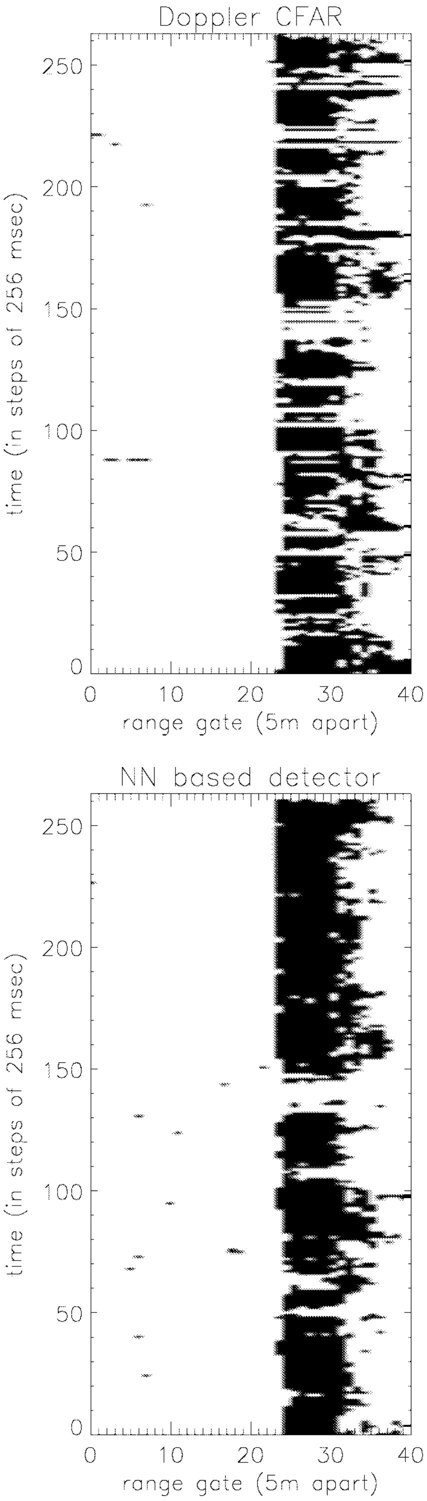


Fig. 7. Postdetection results for the Doppler CFAR and NN receivers.

picture in mind, we see a remarkable improvement in the behavior of the NN receiver in that it fills in the periods of "silence" frequently seen in the detection performance of the conventional Doppler CFAR receiver. This so-called silence is obviously caused by the partial obstruction of the growler (target) by an ocean wave in front of it or the dipping of the growler in a trough. The detection performance displayed in Fig. 7 is indeed quite remarkable. It shows that the NN receiver is able to perform well, even in a situation when the radar returns from the growler are weak. In other words, a "barely visible target is made visible in signal processing terms." The other important observation is the occasional blanking of a signal from the growler (as seen, for example, in the middle of the plot);

in such cases, there is no way any method would be able to detect the target since insofar as the radar is concerned, the target is just not there to be seen.

To describe the detection performance of the receiver of Fig. 4 in traditional quantitative terms, we used a test dataset consisting of a total of 32 292 examples with each example consisting of 256 radar samples, which (as mentioned previously) were completely different from the data used to train the PCA networks and those used to train the MLP's. The results of the tests may be summarized as follows.

1) For a probability of false alarm $P_{FA} > 0.03$, the NN receiver outperformed the Doppler CFAR receiver.
2) For $P_{FA} = 10^{-3}$, the probability of detection (PD) for the two receivers was as follows:

$$P_D = \begin{cases} 0.91, & \text{for the NN receiver} \\ 0.71, & \text{for the Doppler CFAR receiver.} \end{cases}$$

C. Robustness of the Detector

Table 2 tabulates some relevant radar and environmental parameters in the database that was used for the study. The database was made up of four training datasets and nine test datasets. The test datasets had not been previously seen by the receiver, either for self-organized training of the PCA networks or for supervised training of the MLP classifiers. Although the training datasets corresponded to more or less similar environmental conditions, the point to note from this table is that the significant wave height is approximately 1.5 m. However, the test datasets pertained to wave heights varying from 1.5 m–2.6 m. Since the growler protrudes only about 1 m above the water line, the differences in wave heights are significant. Table 2 thus clearly points to the robustness of the NN-based receiver. The robust behavior of the modular detection strategy is attributed directly to the adaptive nature of the receiver, which results from the combined use of self-organized learning and supervised learning for its design.

Table 2

Radar parameters on transmission:

radio frequency:	9.39 GHz
pulse repetition frequency:	2 kHz
pulse duration:	200 ns

Analog-to-digital conversion at the receiver:

sampling rate:	30 MHz
wordlength:	8 bits

Data Sets		Environmental Variables			
Purpose	Number	Significant wave height (m)	Wind speed (kt)	Max. wave height (m)	Peak period, (s)
Training	1	1.59	15	2.6	11.11
	2	1.57	8	2.4	11.11
	3	1.47	18	2.34	11.11
	4	1.46	20	2.27	11.11
Testing	1	1.47	18	2.34	11.11
	2	1.47	21	2.5	11.35
	3	1.46	20	2.27	11.11
	4	2.38	23	3.24	7.69
	5	2.6	20	3.71	9.09
	6	2.23	8	3.47	10.00
	7	2.41	20	3.42	8.00
	8	2.67	18	3.85	7.69
	9	2.42	12	3.61	8.00

VIII. Cost Functions for Supervised Training of the Pattern Classifiers

In the case study on radar detection described in Section VII, we used MLP's for the pattern classifiers in the adaptive receiver of Fig. 4. The MLP's were trained using the BP algorithm. In its traditional form, this algorithm relies on the minimum mean-square error as the optimization criterion. However, the standard criterion for radar detection is the Neyman–Pearson criterion [51]. According to this latter criterion, the probability of detection is maximized subject to a prescribed upper bound imposed on the probability of false alarm [1]–[3]. Unfortunately, minimization of the mean-square error does not guarantee fulfillment of the Neyman–Pearson criterion.

Recognizing that the basic idea of the Neyman–Pearson criterion is to treat the two kinds of error (i.e., missed detection and false alarm) differently, Principe *et al.* [52] have proposed a mixed-norm formulation of the cost functions as follows:

$$\mathcal{E} = \left\{ \begin{array}{ll} \dfrac{1}{N_0} \displaystyle\sum_{t=0}^{N_0-1} |d - y(x(t), \boldsymbol{w})|^{p_0}, & x(t) \in C_0 \\ \dfrac{1}{N_1} \displaystyle\sum_{t=0}^{N_1-1} |d - y(x(t), \boldsymbol{w})|^{p_1} & x(t) \in C_1 \end{array} \right\} \tag{32}$$

where d is the desired response, N_0 is the number of noise-only samples (i.e., hypothesis H_0 is true), and N_1 is the number of target-plus-noise samples (i.e., hypothesis H_1 is true); p_0 and p_1 are the norms for hypotheses H_0 and H_1, respectively; C_0 and C_0 are the corresponding classes of data; $y(x(t), \boldsymbol{w})$ is the receiver output in response to the received signal $x(t)$, parameterized by the weight vector $\boldsymbol{w}$ representing the two pattern classifiers and linear combiner. In order to mimic the Neyman–Pearson criterion with the mixed-norm cost function of (32), two requirements have to be satisfied.

1) Given that hypothesis H_1 is true, the largest deviation of the output $y(x(t), \boldsymbol{w})$ from the desired response d should be minimized so as to set the threshold at the receiver output to as high a level as possible.
2) Given that hypothesis H_0 is true, the influence of large errors should be de-emphasized so that as few of the corresponding output samples exceed the threshold.

These two objectives can be achieved by using the L_∞ norm (i.e., $p_1 = \infty$) for the training examples for which hypothesis H_1 is true, and the L_1 norm (i.e., $p_0 = 1$), or even fractional norms for the training examples for which hypothesis H_0 is true. In Principe *et al.* [52], the necessary modifications to the BP algorithm to work with the cost function of (32) are described.

By training the MLP's in the manner described here, we may expect a further improvement in the performance of the adaptive receiver applied to the radar detection problem discussed in Section VII.

IX. Summary and Discussion

In this paper we have described an adaptive receiver, based on learning, for the detection of a target signal buried in a nonstationary background of unknown statistics. In a fundamental sense the receiver relies on three functional blocks:

1) the transformation of a one-dimensional received signal into 2-D images (maps), whereby the role of time (an essential dimension of learning) is highlighted and the time evolution of the frequency content of the received signal therefore made clearly visible, that is, the received signal is preprocessed such that the target signal and noise components are separated to the best advantage in the "extracted feature space";
2) subspace decomposition for the purpose of dimensionality reduction and therefore efficient learning;
3) pattern classification to pave the way for reliable decision making.

The learning process is self organized in performing the subspace decomposition and supervised in performing the pattern classification. The receiver may therefore be succinctly described as an adaptive detection system based on learning.

The approach taken here is radically different from the classical approach to receiver design. In particular, the need for mathematical modeling of the received signal is eliminated. Instead, successful design of the receiver rests on the availability of a sufficient number of real-life examples representative of the nonstationary environment in which the receiver operates. Part of this database is used to train the receiver, and the remaining part is used to test its performance. The training set is itself split into two parts, one part devoted to design the subspace decomposers and the other part devoted to design the pattern classifiers. Accordingly, the free parameters of the receiver are adjusted in a systematic fashion whereby the information contained in the examples about the environment is extracted and stored in the receiver as parameter values.

In summary, the main virtue of the adaptive two-channel receiver described in this paper is the ability to learn a complex input–output mapping of the environment through a training session. For this learning to be effective, we must have a set of examples representative of the environment: the more representative the examples are, the more robust will the behavior of the receiver be with respect to statistical variations of the environment.

Acknowledgment

The authors wish to thank the anonymous reviewers for their many constructive comments, and they also wish to thank Dr. T. Bhattacharya and B. Currie for reading the final version of the paper and providing comments.

References

[1] H. L. Van Trees, *Detection, Estimation, and Modulation Theory*, pt. 1. New York: Wiley, 1968.

[2] R. N. McDonough and A.D. Whalen, *Detection of Signals in Noise*, 2nd ed. New York: Academic, 1995.

[3] H. V. Poor, *An Introduction to Signal Detection and Estimation*. New York: Springer, 1988.

[4] E. M. Glaser, "Signal detection by adaptive filters," *IRE Trans. Inform. Theory*, vol. IT-7, pp. 87–98, July 1960.

[5] H. Tersranta, S. Urpo, S. Pohjolainen, and E-L. Leskinen, "Measurements of solar oscillations at 37 HGz," in *Proc. Symp. Seismology of the Sun and Sun-like Stars*, Tenerife, Spain, Sept. 1988, pp. 235–239.

[6] L. J. Lanzerotti, D. J. Thomson, and C. G. Maclennan, "Wireless at high altitudes—Environmental effects on space-based assets," *Bell Labs Tech. J.*, vol. 1, pp. 5–9, 1997.

[7] R. P. Lippmann and P. Bakman, "Adaptive neural net preprocessing for signal detection in non-Gaussian noise," *Adv. Neural Inform. Processing Syst.*, vol. 1, pp. 124–132, 1989.

[8] E. Wilson, S. Umesh, and D. W. Tufts, "Multistage neural network structure for transient detection and feature extraction," *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 1993, pp. 489–492.

[9] F.-L. Luo and R. Unbehauen, *Applied Neural Networks for Signal Processing*. Cambridge, UK: Cambridge Univ., 1997.

[10] I. W. Sandberg and L. Xu, "Uniform approximation of multidimensional myopic maps," *IEEE Trans. Circuits Syst.*, vol. 44, pp. 477–485, 1997.

[11] E. Wan, "Temporal backpropagation for FIR neural networks," *Proc. IEEE Int. Joint Conf. Neural Networks*, vol. 1, 1990, pp. 575–580.

[12] L. A. Feldkamp, G. V. Puskorius, and P. C. Moore, "Adaptive behavior for fixed weight networks," *Inform. Sci.*, vol. 98, pp. 217–235, 1997.

[13] Y. LeCun and Y. Bengio, "Convolution networks for images, speech, and time series," in *The Handbook of Brain Theory and Neural Networks*, M. A. Arbib, Ed. Cambridge, MA: MIT, 1995.

[14] S. S. Abeysekera and B. Boashash, "Methods of signal classification using the images produced by the Wigner-Ville distribution," *Patt. Recognit. Lett.*, vol. 12, pp. 717–729, 1991.

[15] D. A. Malhoft, "Neural network approach to the detection problem using joint-time frequency distribution," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, 1990, pp. 2739–2742.

[16] T. K. Bhattacharya and S. Haykin, "Neural network-based radar detection from an ocean environment," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 33, pp. 408–420, Apr. 1997.

[17] S. Haykin and T. K. Bhattacharya, "Modular learning strategy for signal detection in a nonstationary environment," *IEEE Trans. Signal Processing*, vol. 45, pp. 1619–1637, June 1997.

[18] M. Loève, *M. Fonctions aleatoires du second ordre*, Rev. Sc. Paris, t. 84, pp. 195–206, 1946.

[19] ——, *Probability Theory*. New York: Van Nostrand, 1963.

[20] L. Cohen, *Time-Frequency Analysis*. Englewood-Cliffs, NJ: Prentice-Hall, 1995.

[21] F. L. Vernon, III, *Analysis of Data Recorded on the ANZA Seismic Network*, Ph.D. dissertation, Univ. California, San Diego, 1989.

[22] D. J. Thomson, "Spectrum estimation techniques for characterization and development of WT4 waveguide," *Bell Syst. Tech. J.*, vol. 56, pt. I, pp. 1769–1815, pt. II, pp. 1983–2005, 1977.

[23] F. Hlawatsch, "Regularity and unitary of bilinear time-frequency signal representations," *IEEE Trans. Inform. Theory*, vol. 38, pp. 82–94, Jan. 1992.

[24] W. Martin, "Time-frequency analysis of random signals," in *Proc. ICASSP*, 1982, pp. 1325–1328.

[25] P. Flandrin, "On the positivity of the Wigner–Ville spectrum," *Signal Processing*, vol. 11, pp. 187–189, 1986.

[26] D. J. Thomson, "Spectrum estimation and harmonic analysis," *Proc. IEEE*, vol. 70, pp. 1055–1996, Sept. 1982.

[27] D. B. Percival and A. T. Walden, *Spectral Analysis for Physical Applications; Multitaper and Conventional Univariate Techniques*. Cambridge, UK: Cambridge Univ., 1993.

[28] L. I. Tauxe, “Sedimentary records of relative paleointensity of the geomagnetic field; Theory and practice,” *Rev. Geophys.*, vol. 31, pp. 319–354, 1993.

[29] D. J. Thomson, “Time series analysis of Holocene climate data,” *Phil. Trans. R. Soc. Lond. A.*, vol. 330, pp. 601–616, 1990.

[30] ———, “Quadratic-Inverse spectrum estimates; applications to paleoclimatology,” *Phil. Trans. R. Soc. Lond. A.*, vol. 332, pp. 539–597, 1992.

[31] ———, “Nonstationary fluctuations in stationary time-series,” in *Proc. SPIE.*, vol. 2027, 1990, pp. 236–244.

[32] ———, “An overview of multiple-window and quadratic-inverse spectrum estimation methods,” in *Proc. ICASSP.*, vol. 6, 1994, pp. 185–194.

[33] A. Papoulis, “A new algorithm in spectral analysis and band-limited extrapolation,” *IEEE Trans. Circuits and Systems*, vol. CAS-22, pp. 735–742, 1975.

[34] F. Gori and C. Palma, “On the eigenvalues of sinc^2 kernel,” *J. Phys. A: Math. Gen.*, vol. 8, pp. 1709–1719, 1975.

[35] D. J. Thomson, “Dependence of global temperatures on atmospheric CO_2 and solar irradiance,” *Proc. Natl. Acad. Sci. USA*, vol. 94, pp. 8370–8377, 1977.

[36] R. J. Mellors, F. Vernon, and D. Thomson, “Detection of dispersive signals using multi-taper dual-frequency coherence,” in *Proc. 18th Seismic Research Symp. Monitoring a Comprehensive Test Ban Treaty*, Annapolis, MD, 1996, pp. 745–753.

[37] R. J. Mellors, F. L. Vernon, and D. J. Thomson, “Detection of dispersive signals using multitaper dual-frequency coherence,” *Geophysical J. Int.*, to be published.

[38] R. Schild and D. J. Thomson, “The Q0957 + 561 time delay,” in *Quasar Structure and Microlensing* (Astronomical Time Series), D. Maoz *et al.*, Eds. Norwell, MA: Kluwer, 1997, pp. 73–84.

[39] H. L. Hurd, “Spectral coherence of nonstationary and transient stochastic processes,” in *Proc. 4th IEEE ASSP Workshop Spectrum Estimation and Modeling*, Minneapolis, MN, 1988, pp. 387–390.

[40] N. L. Gerr and J. C. Allen, “The generalized spectrum and spectral coherence of a harmonizable time series,” *Digital Signal Processing*, vol. 4, pp. 222–238, 1994.

[41] G. C. Carter, Ed., *Coherence and Time Delay Estimation*. New York: IEEE, 1993.

[42] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. New York: Wiley, 1973.

[43] K. Fukunaga, *Statistical Pattern Recognition*, 2nd ed. New York: Academic, 1990.

[44] S. Haykin, “Neural networks expand SP’s horizons,” *IEEE Signal Processing Mag.*, vol. 13, pp. 24–29, Feb. 1996.

[45] ———, *Neural Networks: A Comprehensive Foundation*. New York: Macmillan, 1994.

[46] N. Suga, “Cortical computational maps for auditory imaging,” *Neural Networks*, vol. 3, pp. 3–21, 1990.

[47] J. A. Simmons, “Time-frequency transforms and images of targets in the sonary of bats,” *NEC Res. Inst.*, Princeton, NJ, 1991.

[48] S. Haykin, C. Krasnor, T. Nohara, B. Currie, and D. Hamburger, “A coherent dual-polarized radar for studying the ocean environment,” *IEEE Trans. Geosci. Remote Sensing*, vol. 29, p. 1890–18191, 1991.

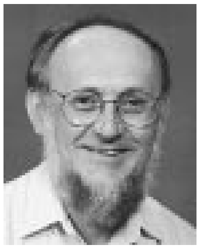
[49] T. D. Sanger, “Optimal unsupervised learning in a single-layer linear feedforward network,” *Neural Networks*, vol. 1, pp. 459–473, 1989.

[50] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, “Learning representations by back-propagating errors,” *Nature*, vol. 323, pp. 533–536.

[51] J. Newman and E. S. Pearson, “On the problem of the most efficient tests as statistical hypotheses,” *Phil. Trans. Roy. Soc., London, Series A*, vol. 231, pp. 289–337, 1933.

[52] J. C. Principe, M. Kim, and J. W. Fisher, III, “Target discrimination in synthetic aperture radar (SAR) using artificial neural networks,” *IEEE Trans. Image Processing*, to be published.

Simon Haykin (Fellow, IEEE), for a photograph and biography, see this issue, p. 2120.



David J. Thomson (Fellow, IEEE) was born in Saint John, N.B., Canada in 1942. He received the B.Sc. degree (Honors) in mathematics from Acadia University, Wolfville, N.S., Canada, in 1965 and the M.S. and Ph.D. degrees in electrical engineering from the Polytechnic Institute of Brooklyn, Brooklyn, NY, in 1967 and 1971, respectively.

Since 1965 he has been a Member of the Technical Staff at Bell Laboratories and has worked on multipair and coaxial cable development and the WT4 Millimeter Waveguide System. In the Advanced Mobile Phone Service project he was responsible for the circuit design of and software for a microprocessor-controlled modem for Rayleigh fading channels. He has been an Adjunct Professor in the Graduate Department of Scripps Institution of Oceanography and at the Neurological Institute of Columbia University. He has taught statistical inference at Princeton University, time series at Stanford University, and he gave the Houghton Lectures at Massachusetts Institute of Technology. He has been a Green Scholar at Scripp’s Institution of Oceanography and the Steinbeck Visiting Scholar at Woods Hole Oceanographic Institution. Presently he is a Distinguished Member of Technical Staff in the Communications Analysis Research Department at Bell Laboratories. In addition to spectrum estimation and communications theory, his present research interests include digital signal processing, robust statistics, phase tracking and time delay problems, modulation theory, circuit design, seismology, paleoclimatology, global climate, and gravitational lensing. He has written over 75 papers and has 15 patents, most dealing with his primary interests of signal processing and time series analysis.

Dr. Thomson is a member of the American Geophysical Union, the American Statistical Association, and the Institute of Mathematical Statistics. He is a Chartered Statistician and a Fellow of the Royal Statistical Society. He is Chairman of Commission C of USNC-URSI, and he is an Associate Editor for *Radio Science*. He was also an Associate Editor of IEEE Transactions on Information Theory.