XX. PROCESSING AND TRANSMISSION OF INFORMATION*

Academic and Research Staff

Prof. P. Elias
Prof. R. M. Gallager
Prof. M. M. Goutmann
Prof. E. V. Hoversten
Prof. D. A. Huffman
Prof. R. S. Kennedy
Prof. J. L. Massey
Prof. E. Mortenson
Prof. C. E. Shannon
Prof. R. N. Spann

Graduate Students

D. S. Arnstein
E. A. Bucher
D. Chase
R. L. Greenspan
H. M. Heggestad
J. A. Heller
M. Khanna
Jane W-S. Liu
J. Max
J. C. Moldon
J. T. Pinkston III
E. M. Portner, Jr.
J. S. Richters
D. A. Wright

A. OPTICAL IMAGE ESTIMATION IN A TURBULENT ATMOSPHERE

A major performance limitation of ground-based telescopes is now imposed by the effects of turbulence in the atmosphere. Because of random variations in index of refraction, such instruments are seldom able to achieve useful resolutions better than one second of arc. In this report we shall first describe the statistical nature of this problem, and review the limitations encountered with present image-recording techniques such as photographs taken at the focal plane of a telescope. We shall then describe some of the results of our investigations indicating that a significant improvement in resolution may be possible if one employs appropriate data collection and processing techniques.

1. Problem Model

The geometry for the problem is shown in Fig. XX-1. In the absence of a turbulent atmosphere, we suppose that the complex amplitude of the received field $E_a(\underline{x}, t)$ can be written

$$E_a(\underline{x}, t) = \int d\underline{\xi}\ E_s(\underline{\xi}, t)\ e^{+j2\pi \frac{\underline{\xi} \cdot \underline{x}}{\lambda}},$$

where

$E_s(\underline{\xi}, t)$ is the complex amplitude of the electric field radiated from a point ξ on the object: for simplicity, we treat this field as a scalar function.

*This work was supported principally by the National Aeronautics and Space Administration (Grant NsG-334; and in part by the Joint Services Electronics Programs (U.S. Army, U.S. Navy, and U.S. Air Force) under Contract DA 28-043-AMC-02536(E).

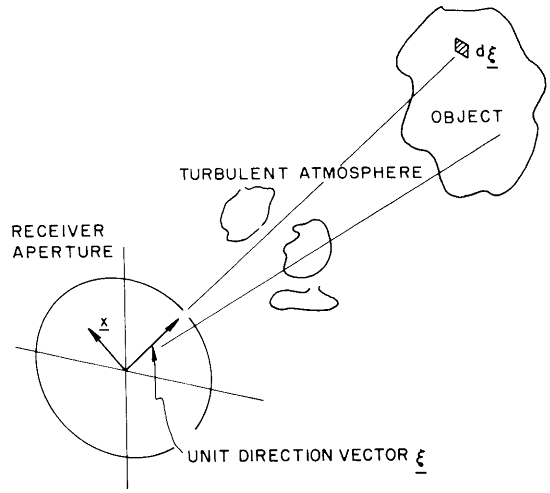

Fig. XX-1. Geometry for the problem.

$\underline{\xi}$ is the coordinate vector locating points on the object, measured as a direction vector, the object being situated at a sufficient distance that the incoming waves are essentially planar.

$\underline{x}$ is the spatial coordinate in the plane of the receiving aperture.

λ is the wavelength of the light, which, for simplicity, we have assumed to be almost monochromatic.

We modify this expression to account for turbulence by introducing a random time and spatially varying complex multiplicative factor. Thus, with turbulence present,

$$E_a(\underline{x}, t) = \int d\underline{\xi}\, E_s(\xi, t)\, e^{j2\pi \frac{\underline{\xi} \cdot \underline{x}}{\lambda}}\, e^{\alpha(\underline{\xi}, \underline{x}, t) - j\beta(\underline{\xi}, \underline{x}, t)}.$$

The terms $\alpha(\underline{\xi}, \underline{x}, t)$ and $\beta(\underline{\xi}, \underline{x}, t)$ are the log amplitude and phase disturbances introduced by the atmosphere. When this expression is introduced into the wave equation for an inhomogeneous medium the solutions for both α and β appear as functions of the index of refraction integrated over the path. Since the integration encloses many statistically independent pieces of the atmosphere, it is generally agreed that α and β should behave as jointly Gaussian random processes.[1] In general, these processes will depend upon the path direction $\underline{\xi}$, as well as on $\underline{x}$ and t. In the interest of simplicity in outlining the problem, we shall invoke the isoplanatic condition that there is no $\underline{\xi}$ dependence.

It is more convenient to view the received phenomena in terms of a mutual coherence function, rather than the actual instantaneous field values. We define this mutual coherence function as

$$\Gamma(\underline{x}, \underline{x}', t) \triangleq \overline{E_a(\underline{x}, t)\, E_a^*(\underline{x}', t)},$$

where the bar refers to an ensemble average over the source but not over the atmospheric fluctuations. Furthermore, although we have restricted the light to be narrowband in order to simplify the expression for the received field, this bandwidth can be many, many times greater than the bandwidths of the turbulence-induced fluctuations. Suppose that one actually multiplies together values of the field at points $\underline{x}$ and $\underline{x}'$ and time averages over an interval that is short in comparison with the time scale of atmospheric fluctuations. This interval can still be sufficiently long that with overwhelming probability the measured mutual coherence function will be the same as the ensemble average defined above.

The mutual coherence function on the aperture can be expressed in terms of a source mutual coherence function by

$$\Gamma_a(\underline{x}, \underline{x}', t) = e^{\alpha(\underline{x}, t)+\alpha(\underline{x}', t)-j[\beta(\underline{x}, t)-\beta(\underline{x}', t)]} \iint d\underline{\xi} d\underline{\xi}' \; \Gamma_s(\underline{\xi}, \underline{\xi}', t) \; e^{+j2\pi[\underline{\xi}\cdot\underline{x}-\underline{\xi}'\cdot\underline{x}']}$$

in which time stationarity has been assumed for the statistics of the source radiation.

A further reasonable assumption for naturally illuminated objects is spatial incoherence. That is, $\Gamma_s(\underline{\xi}, \underline{\xi}', t)$ has the form

$$\Gamma_s(\underline{\xi}, \underline{\xi}', t) = \phi(\underline{\xi}) \; \delta(\underline{\xi}-\underline{\xi}')$$

such that

$$\Gamma_a(\underline{x}, \underline{x}', t) = e^{\alpha(\underline{x}, t)+\alpha(\underline{x}', t)-j[\beta(\underline{x}, t)-\beta(\underline{x}', t)]} \int d\underline{\xi} \; \phi(\underline{\xi}) \; e^{+j2\pi \frac{\underline{\xi}\cdot(\underline{x}-\underline{x}')}{\lambda}}.$$

The function $\phi(\underline{\xi})$ describes the distribution of light intensity over the object. This function, of course, is the one that we seek to determine.

2. Conventional Photography

The field $E_f(\underline{u}, t)$ at the focal plane of a lens can be approximated by

$$E_f(\underline{u}, t) = \frac{1}{\lambda L} \int d\underline{x} \; w(\underline{x}) \; E_a(\underline{x}, t) \; e^{+j2\pi \frac{\underline{x}\cdot\underline{u}}{\lambda L}}.$$

The focal length of the lens is L, and $w(\underline{x})$ is a pupil function having value 1 for $\underline{x}$ within the exposed aperture, and 0 outside.

The average intensity of this focal plane is, then,

$$I_f(u, t) = \overline{|E_f(u, t)|^2}$$

$$= \left(\frac{1}{\lambda L}\right)^2 \int d\underline{x} \int d\underline{x}' \; w(\underline{x}) \; w(\underline{x}') \; \Gamma_a(\underline{x}, x', t) \; e^{+j2\pi \frac{u\cdot(\underline{x}-\underline{x}')}{\lambda L}}.$$

When one is concerned with angular resolution, it is convenient to work with spatial Fourier transforms. Thus we define as transforms of the object and image

$$\Theta(\underline{k}) = \int d(\underline{\xi})\ \phi(\underline{\xi})\ e^{+j2\pi\underline{\xi}\cdot\underline{k}}$$

and

$$F(\underline{k},t) = \int d\underline{u}\ I(\underline{u},t)\ e^{-j2\pi\frac{\underline{u}\cdot\underline{k}}{L}}.$$

We then have

$$\Gamma_a(\underline{x},\underline{x}',t) = \Theta\left(\frac{\underline{x}-\underline{x}'}{\lambda}\right) e^{\alpha(\underline{x},t)+\alpha(\underline{x}',t)-j[\beta(\underline{x},t)-\beta(\underline{x}',t)]}$$

$$I_f(u,t) = \left(\frac{1}{\lambda L}\right)^2 \int d\underline{x} \int d\underline{x}'\ w(\underline{x})\ w(\underline{x}')\ e^{+j2\pi\frac{\underline{u}\cdot(\underline{x}-\underline{x}')}{\lambda L}}\ e^{\alpha(\underline{x},t)+\alpha(\underline{x}',t)-j[\beta(\underline{x},t)-\beta(\underline{x}',t)]}\,\Theta\left(\frac{\underline{x}-\underline{x}'}{\lambda}\right)$$

and

$$F(\underline{k},t) = H(\underline{k})\ \Theta(\underline{k}), \tag{1}$$

where

$$H(\underline{k}) = \int d\underline{x}\ w(\underline{x})\ w(\underline{x}+\lambda\underline{k})\ e^{\alpha(x,t)+\alpha(\underline{x}+\lambda\underline{k},t)+j[\beta(\underline{x},t)-\beta(\underline{x}+\lambda\underline{k},t)]}.$$

This last function, $H(\underline{k})$, is called the instantaneous modulation transfer function (MTF) of the system.

A conventional photograph records the integral of $I_f(\underline{u},t)$ with respect to t over the exposure time. When the exposure time is quite long compared with the atmosphere's correlation time, the photograph that would result if $H(\underline{k})$ were replaced by its ensemble average value, under the assumptions that α and β are Gaussian, stationary, and homogeneous, is the average $\overline{H}(\underline{k})$ that may be expressed as

$$\overline{H}(\underline{k}) = H_o(\underline{k})\ H_1(\underline{k}).$$

The first term, $H_o(\underline{k}) = \int w(\underline{x})\ w(\underline{x}+\lambda\underline{k})\ d\underline{x}$, is just the familiar MTF associated with a lens of finite size. The second term, $H_1(\underline{k}) = e^{-\frac{1}{2}\mathscr{D}(|\lambda\underline{k}|)}$, is the contribution from the turbulence. $\mathscr{D}$ is the wave structure function and is defined as

$$\mathscr{D}(r) \overset{\Delta}{=} \overline{[\alpha(\underline{x},t)-\alpha(\underline{x}+\underline{r},t)]^2 + [\beta(\underline{x},t)-\beta(\underline{x}+\underline{r},t)]^2}.$$

Most of $\mathscr{D}$ is made up from the phase disturbances, and both theoretical and experimental investigations[1] indicate that $\mathscr{D}(r)$ varies as $r^{5/3}$. We choose to use Frieds[2] form

$$\mathscr{D}(r) = 6.88\left(\frac{r}{r_o}\right)^{5/3}.$$

The coefficient r_o can be considered as a characteristic phase correlation distance on the aperture. Thus we have

$$H_1(\underline{k}) = e^{-3.44\left(\frac{\lambda|k|}{r_o}\right)^{5/3}}.$$

It is evident that with a long exposure the atmosphere attenuates very sharply frequencies beyond what we can call an upper cutoff frequency given by r_o/λ.

It has been observed[4] that the resolution can be improved simply by making a very long exposure and then filtering with the inverse of the average MTF. Such techniques are not very efficient because the sample average MTF will differ from the ensemble average in any finite time interval. With the extremely rapid roll-off of the average MTF, we expect the inverting operation to tremendously enhance such residual errors at high spatial frequencies. Indeed, the mean-square error, $\overline{|\epsilon(k)|^2}$, as a function of spatial frequency is

$$\overline{\left|\frac{\epsilon(\underline{k})}{\Theta(\underline{k})}\right|^2} = \frac{1}{N}\left[e^{\mathscr{D}(\lambda\underline{k})+4\overline{a(\underline{x},t)a(\underline{x}+\lambda\underline{k},t)}} - 1\right], \tag{2}$$

where N is the number of independent realizations or "looks" at the channel. Thus if one must keep the error below some given value, the number of independent "looks" grows exponentially with spatial frequency.

A more popular technique for improving resolution is the use of very fast exposures which "freeze" the turbulent motions of the atmosphere. Although time smearing is eliminated, spatial averaging of the disturbances across the aperture still results in a smearing effect. With a small aperture, however, much of the phase distortion can be ascribed to a "tilting" of the wave fronts which results in a gross motion of the focal plane image. For systems that can track this motion, Fried[3] has calculated that the optimum aperture diameter is $\sim 3r_o$, and that the upper cutoff frequency of the resultant average MTF is increased by approximately a factor of 3.

3. Proposed Technique

Imagine for the moment that instead of recording the intensity in the focal plane, we make point measurements of the mutual coherence function over the aperture. Then we would have a set of measurements of the form

$$\Gamma(\underline{x}_i, \underline{x}'_i, t_i) = \Theta\left(\frac{\underline{x}_i - \underline{x}'_i}{\lambda}\right) e^{a(\underline{x}_i, t_i) + a(\underline{x}'_i, t_i) - j[\beta(\underline{x}_i, t_i) - \beta(\underline{x}'_i, t_i)]}.$$

These measurements constitute a sufficient statistic when detector noise can be ignored and when, also, the field radiated from the object is a Gaussian random process that is wideband relative to the bandwidth of the atmospheric fluctuations.

Taking the logarithm of these complex numbers would yield

$$\log \Gamma(\underline{x}_i, \underline{x}'_i, t_i) = \log \Theta\left(\frac{\underline{x}_i - \underline{x}'_i}{\lambda}\right) + a(\underline{x}_i, t_i) + a(\underline{x}'_i, t_i) - j[\beta(\underline{x}_i, t_i) - \beta(\underline{x}'_i, t_i)].$$

To obtain a maximum likelihood estimate of the function $\phi(\xi)$, we can ask for the maximum-likelihood estimate of log $\Theta(\underline{k})$. This estimate follows immediately as

$$\underset{\text{likelihood}}{\max}\ \{\log \Theta(\underline{k})\} = \left\langle \log \Gamma(\underline{x}_i, \underline{x}_i + \lambda\underline{k}, t_i)\right\rangle - 2\,\overline{a(\underline{x}, t)},$$

where brackets denote a sample average of the log of the observed mutual coherence function. If the only disturbances that the system needs to contend with are the statistical fluctuations in a and β, the mean-square error in estimating the component at spatial frequency $\underline{k}$ is

$$\overline{\left|\frac{\epsilon(\underline{k})}{\Theta(\underline{k})}\right|^2} = e^{\frac{2}{N}\overline{[a(\underline{x}, t) + a(\underline{x}+\lambda\underline{k}, t)]^2}} - 2\, e^{\frac{2}{N}\overline{a(\underline{x}t)^2} - \frac{1}{2N}\mathcal{D}(\lambda\underline{x})} + 1.$$

When N, the number of independent "looks" is large enough that $\overline{\left|\frac{\epsilon(\underline{k})}{\Theta(\underline{k})}\right|^2} \lesssim 0.1$, this expression is closely approximated by the simpler form

$$\overline{\left|\frac{\epsilon(\underline{k})}{\Theta(\underline{k})}\right|^2} \approx \frac{1}{N}[\mathcal{D}(\lambda\underline{k}) + 4\overline{a(\underline{x}, t)a(\underline{x}+\lambda\underline{k}, t)}] \tag{3}$$

A comparison of this expression with that for inverse filtering (Eq. 2) is quite startling. With inverse filtering the error grows exponentially with spatial frequency, with this technique the error grows only algebraically. Thus such a technique, if physically feasible, appears to give promise of significant improvement over conventional techniques in resolution capability.

Let us now see just what is involved in measuring the mutual coherence function, and consider what additional disturbances are inevitably introduced. At long wavelengths (microwaves), such correlation measurements could be made by using two small detectors in the aperture. At optical frequencies such direct methods are precluded, because of the extremely small scale of the field pattern, and we must use an indirect method.

If the pupil function, $w(\underline{x})$, consists of two very small holes of area A centered at points $\underline{x}_i$ and $\underline{x}_i + \lambda\underline{k}$, we find, from Eq. 1, that

$$F(\underline{k}, t) = A\ \Theta(\underline{k})\ e^{\alpha(\underline{x}, t)-\alpha(\underline{x}+\lambda\underline{k}, t)+j[\beta(\underline{x}, t)-\beta(\underline{x}+\lambda\underline{k}, t)]},$$

provided that A is smaller than the coherence area for the phase and amplitude disturbances. Thus by suitably shading the aperture of the lens, recording the focal plane intensity with a fast exposure, and finally correlating the result with $\exp\left[-j2\pi\frac{\underline{k}\cdot\underline{u}}{L}\right]$, one obtains the real and imaginary components of $\Gamma(\underline{x}_i, \underline{x}_i+\lambda\underline{k}, t_i)$.

It is in converting this measurement into phase and log amplitude form (taking the logarithm) that problems arise which will ultimately impose a limitation on the utility of the log coherence function averaging technique. In any recording device, for example, photographic film, the effect of some unavoidable background noise must be considered. If this noise is small, the observed phase differs only slightly from the true phase of Γ, and one would expect that this will become unimportant in the subsequent averaging operation. When the noise is large, however, dominating the signal, the measured phase will be completely unrelated to the actual phase. In the succeeding phase measurements, even if the noise is small, it is possible to have gained or lost 2π radians in the absolute value of the phase. If such gross phase errors occur frequently enough, the technique will yield a very poor estimate of the phase of the spatial Fourier transform of the object.

This limitation of the technique is, of course, the familiar anomaly problem common to all nonlinear modulation schemes (for example, the "click" problem of phase and frequency modulation systems). It becomes important when the signal-to-noise ratio drops below a critical threshold value, beyond which the system performance deteriorates drastically. Since the amount of signal power incident on the recording device is proportional to the hole area A, we want to make A as large as possible without incurring too much of the undesirable spatial averaging of the phase fluctuations. The optimum hole diameter is probably $\sim r_o$, the characteristic phase correlation distance. Similarly, one wants to integrate this signal power over one correlation time constant of the disturbances.

We have presented a rather simple processor illustrating what we believe are the fundamental problems to be encountered in image estimation. If the source is very bright, we are well above threshold and the simple technique will yield a mean-square error of Eq. 3. When operating with dimmer objects, good performance may still be attainable by processing the mutual coherence function with operators that exploit more of the statistical dependences in the received process. One such dependence is the correlation of the disturbances at different spatial frequencies within a coherent spatial bandwidth (given approximately by λ/r_o).

4. Further Considerations

Two rather limiting, but expedient, assumptions were made in the preceding discussion. The most important of these is the isoplanatic assumption, that the phase and amplitude disturbances induced by the atmosphere are the same for all direction angles. Often this is not the case, but we can suppose that there is some characteristic angular spread, ξ_o, over which the disturbances appear to be essentially the same. We should then be able to independently apply our processing technique to each of these isoplanatic patches. In order to separate these patches in the focal plane, however, it is necessary that $\frac{\lambda}{\xi_o r_o} < 1$.

We have previously noted that λ/r_o is the coherent spatial bandwidth of the channel. Similarly, $1/\xi_o$ can be interpreted as the spatial frequency dispersion bandwidth. Thus we borrow a term from the field of scatter communication links and denote the condition above as the underspread channel and its converse as the overspread channel.

The other major assumption was that the light was very narrow-band. In order to find the extent of this requirement, we imagine the light to be composed of a spectrum of very narrow processes and limited to a wavelength range $\lambda_{min} < \lambda < \lambda_{max}$. According to a ray theory analysis[5] of wave propagation through turbulence, the ray path delay time $\tau(x,t)$ is a random variable independent of wavelength (provided the index of refraction is the same for all wavelengths). The wave phase, as a function of wavelength, is then

$$\beta_\lambda(\underline{x},t) = \frac{2\pi c}{\lambda}\,\tau(\underline{x},t),$$

where c is the speed of light. Equation 1 should now be integrated over λ. It is possible that this integration can also lead to image smearing, unless the phase differences at points $\underline{x}$ and $\underline{x}'$ are about the same for all wavelengths. The expected squared value of this range of differences is

$$\overline{|\Delta\beta|^2} = \overline{\left[2\pi c\,\frac{\lambda_{max}-\lambda_{min}}{\lambda_{max}\lambda_{min}}\,[\tau(\underline{x},t)-\tau(\underline{x}',t)]\right]^2}$$

$$\approx \left(\frac{\Delta\lambda}{\lambda_o}\right)^2 \mathscr{D}_{\lambda_o}[|\underline{x}-\underline{x}'|],$$

where

$$\Delta\lambda = \lambda_{max} - \lambda_{min}$$

$$\lambda_o = \sqrt{\lambda_{max}\lambda_{min}} \qquad \text{(geometric mean).}$$

A reasonable requirement is to demand that $\overline{|\Delta\beta|^2}$ be less than ~7, in order to avoid color smearing. Thus we want

$$\left[\frac{|x-x'|}{r_o}\right]^{5/3} < \left(\frac{\lambda_o}{\Delta\lambda}\right)^2.$$

It is interesting to note that such filtering is implicit in a measurement in which an aperture mask is used with two holes of diameter r'_o, centered at $\underline{x}$ and $\underline{x}'$. The only optical wavelengths that will contribute to spatial frequency component $\underline{k}$, in the image, fall within $\lambda'_{max} = \frac{|\underline{x}-\underline{x}'| + r'_o}{k}$ and $\lambda'_{min} = \frac{|\underline{x}-\underline{x}'| - r'_o}{k}$, where $\underline{k} = \frac{|\underline{x}-\underline{x}'|}{\lambda_o}$; and hence $\frac{\Delta\lambda'}{\lambda_o} = \frac{2r'_o}{|x-x'|}$.

This implicit bandlimiting will then satisfy our requirements, provided

$$r'_o \lesssim \frac{1}{2} r_o \left[\frac{|\underline{x}-\underline{x}'|}{r_o}\right]^{1/6};$$

and thus the hole diameter need not be too much smaller than the optimum diameter, r_o.

J. C. Moldon

References

1. V. I. Tatarski, Wave Propagation in a Turbulent Medium (McGraw-Hill Book Company, New York, 1961).

2. D. L. Fried, "Statistics of Geometric Representation of Wavefront Distortion," J. Opt. Soc. Am. 55, 11 (1965).

3. D. L. Fried, "Optical Resolution through a Randomly Inhomogeneous Medium for Very Long and Very Short Exposures," J. Opt. Soc. Am. 56, 10 (1966).

4. J. L. Harris, "Image Evaluation and Restoration," J. Opt. Soc. Am. 56, 5 (1966).

5. R. S. Kennedy and E. V. Hoversten, "Optical Propagation through a Turbulent Atmosphere," Quarterly Progress Report No. 84, Research Laboratory of Electronics, M.I.T., January 15, 1967, pp. 212-224.

B. QUANTUM COMMUNICATION SYSTEMS

We are concerned with the detection of a discrete set of M signals that are transmitted through a channel disturbed by chaotic thermal noise. Quantum limitations on the precision of measurements, as well as quantization of the radiation fields, will be taken into account. To be more specific, we would like to determine, from the statistics of the noise field, the types of measurements that should be made by the optimum receiver, and bounds on the resulting probability of detection error.

The model of the communication systems that we shall study is shown in Fig. XX-2.

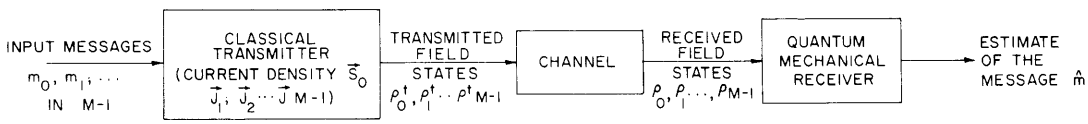

Fig. XX-2. Quantum communication system model.

The M input messages to the transmitter, $m_0, m_1, \dots m_{M-1}$, are assumed to be equiprobable. A message, m_s, is represented by the state of the transmitted electromagnetic wave given by the density operator, ρ_s^t. Therefore, depending on the input message to the transmitter, the transmitted electromagnetic field is in one of M states given by the density operators $\rho_0^t, \rho_1^t, \dots, \rho_{M-1}^t$. Because of inverse square law loss, as well as the disturbance of additive background noise in the channel, the state of electromagnetic field at the output of the channel (the received field) will be one of the M states given by the density operators $\rho_0, \rho_1, \dots, \rho_{M-1}$, which are no longer the same as those specifying the transmitted states $\rho_0^t, \rho_1^t, \dots, \rho_{M-1}^t$.

We wish to establish a strategy of making an estimate, m, of the transmitted message, m, such that the probability of error, $P(m \neq m)$, is minimal. In the quantum communication system, the received field contains all of the relevant data. The receiver makes a set of measurements of the field and, based on results of the measurements, estimates which one of the set of density operators best describes the field and therefore which of the messages is transmitted. As will become apparent in this report, the quantum formulation will not enable us to determine the operations performed by the optimum receiver without introducing some constraints on these operations. Hereafter, instead of the optimum receiver, we shall discuss the problem of finding the receiver that yields the smallest probability of error among the receivers in a given class.

1. Signal Fields

Physically, the transmitter in Fig. XX-2 can be considered as a classical current source of the electromagnetic field, for example, a current density on the transmitting aperture. When the input message is m_s, the current density in the transmitting aperture, $\vec{J}_s(\underline{r}, t)$, generates the transmitted field in a state given by the density operator ρ_s^t. By saying that the current density is classical, we mean that its reaction with the process of radiation is either negligible, or at least predictable, in principle. Such a model for the source is justified, since the power of the source is usually sufficiently high that it can be described classically. In many practical systems, we can further simplify the problem by approximating the transmitter by a point source. Then, $\vec{J}_s(\underline{r}, t)$ can be written as $\delta(\underline{r})\, I_s(t)\, \vec{e}_t$.

The vector potential, $\vec{A}(\underline{r}, t)$, of the electromagnetic field of finite duration generated by the transmitter can be expanded in terms of an appropriate set of vector normal mode functions $\{\vec{U}_k(\underline{r})\}$. That is,

$$\vec{A}(\underline{r}, t) = c \sum_k \sqrt{\frac{\hbar}{2\omega_k}} \left(a_k \vec{U}_k(r)\, e^{-i\omega_k t} + a_k^+ \vec{U}_k^*(r)\, e^{i\omega_k t} \right), \tag{1}$$

where a_k and a_k^+, The annihilation and creation operators of the k^{th} mode, satisfy the commutation relations

$$\left[a_k^+, a_j^+\right] = [a_k, a_j] = 0; \qquad \text{for all } k \text{ and } j$$

$$\left[a_k, a_j^+\right] = \delta_{kj}. \tag{2}$$

For simplicity, we assume that the geometry of the transmitter aperture is such that only one spatial mode for each frequency is transmitted.

From relations (2), it is apparent that the dynamical variables of different modes can be observed independently of one another. It follows that the observation of most of the dynamical variables of the entire field, for example, the Hamiltonian H, $\vec{E}(\underline{r}, t)$, $\vec{H}(\underline{r}, t)$, etc., can be reduced to the observations of the dynamic variables of the individual modes separately. Therefore, we shall begin by limiting ourselves to those receiving systems that indeed observe the individual modes of the field separately.

The receiving system is assumed to consist of a set of noninteracting cavities, located at a distance d from the transmitter. Each of these cavities has only one dominant mode, and there is exactly one cavity at each one of the normal mode frequencies $\omega_1, \omega_2, \ldots \omega_k, \ldots$ of the field. These cavities are initially empty and are exposed to the signal source by opening their apertures when the signal is expected to arrive. At the end of this time interval, the apertures are closed by shutters.

It can be shown that inside the cavity with dominant frequency $\omega_k (k=1, 2, \ldots)$, the electromagnetic field generated by the current source $\vec{J}_s(\underline{r}, t)$ corresponding to the message m_s is in a coherent state given by the state vector $|\alpha_k^s(t)\rangle_k$ at time t, where $a_k^s(t)$ is given by

$$a_k^s(t) = i \frac{\ell}{\sqrt{V_r}} \frac{1}{4\pi dc} \sqrt{\frac{\omega_k}{\hbar}} \int_{-\infty}^{t} dt' \; I_s\left(t' - \frac{d}{c}\right) e^{-i\omega_k t'} \tag{3}$$

A state of the field is said to be a coherent state when the state vector $|\{a_k\}, t\rangle$ is a right eigenvector of the operator $\vec{E}(r,t) = i \sum_k \sqrt{\frac{\hbar\omega_k}{2}} \left(a_k \vec{U}_k(r) e^{-i\omega_k t}\right)$. This implies that the k^{th} normal mode is in the coherent state $|a_k\rangle_k$, which satisfies the relation $a_k|a_k\rangle_k = a_k|a_k\rangle_k$; for all k. Hence the coherent state $|\{a_k\}\rangle$ of the entire field is just the direct product $\pi_k |a_k\rangle_k$.

Expression (3) is valid when the receiving aperture of the cavity is polarized parallel to the transmitting aperture and the loss of the cavity is negligible. In this equation, ℓ is the linear dimension of the receiving aperture, and V_r is the volume of the cavity. The total electromagnetic field excited by the signal source alone is in the state given by the vector $\pi_k |a_k^s(t)\rangle_k$. This state is also given by the density operator ρ_s^t, which in the P-representation is just

$$\rho_s^t = \int \pi_k \, \delta^2\left(\beta_k - a_k^s(t)\right) |\beta_k\rangle_k \, {}_k\langle\beta_k| \, d^2\beta_k, \tag{4}$$

where $a_k^s(t)$ is given by Eq. 3.

2. Noise Fields

Four types of noise are present in a quantum-mechanical communication channel: the source quantization noise, the partition noise, the quantum noise, and the additive background. The source quantization noise need not be considered here, since we have assumed that the signal source is a classical one. The partition noise is associated with the inverse square law attenuation present in many practical communication systems. Such loss has, in fact, already been accounted for in the model of the receiving system consisting of a set of resonant cavities. Therefore, there is no need to consider this loss separately. The "quantum noise" is usually introduced to take into account the uncertainties in the quantities measured by the receiving system which, according to quantum theory, cannot make precise measurements of all of the dynamical variables of the signal field. This type of noise will emerge naturally from the formulation of the problem.

By additive noise, we mean the chaotic background radiation which is also present in the receiving cavities, but is not excited by the signal source. This noise field can be considered to be generated by a completely incoherent source modeled as a large number of independent stationary sources. Glauber[2] has shown that at thermal equilibrium, the noise field is in a completely incoherent state given by the density operator, which in the P-representation has the weight function

$$P(\{\alpha_k\}) = \prod_k \frac{1}{\pi\langle n_k\rangle} \exp\left[-\frac{|\alpha_k|^2}{\langle n_k\rangle}\right]. \tag{5}$$

$\langle n_k\rangle$ in this equation is the average number of photons in the k^{th} mode. At temperature T, it is given by

$$n_k = \frac{1}{\exp\left(\frac{\hbar\omega_k}{kT}\right) - 1}. \tag{6}$$

The received field, the field excited by the signal source and the noise sources, is just a superposition of the transmitted field and the noise field. When the input message is m_s, the received field is in the state specified by the density operator

$$\rho_s = \int P_s\,(\{\beta_k\})\,|\{\beta_k\}\rangle\langle\{\beta_k\}|\,\prod_k d^2\beta_k \tag{7}$$

$$P_s(\{\beta_k\}) = \prod_k \frac{1}{\pi\langle n_k\rangle}\,\exp - \frac{|\beta_k - \alpha_k^s(t)|^2}{n_k}.$$

We shall confine the discussion in this report to the simplest signal set, the binary on-off set. The results obtained can be easily generalized to other sets of signals with finite and discrete spectra. The equally likely binary signals we consider are defined by the correspondences

$$m = m_0 \longleftrightarrow I(t) = 0$$

$$m = m_1 \longleftrightarrow I(t) = I_j \cos(\omega_j t+\phi),$$

where ω_j is some fixed, but arbitrary, mode frequency.

It can be shown that the individual modes of the signal field inside the cavities in the absence of the noise are uncorrelated and that the noiseless signal field is in a state $|\{\alpha_k^s\}\rangle$, where the complex amplitudes α_k^s are

$$\alpha_k^s = \sigma_{ks}\exp(-i\theta_{ks}), \tag{8}$$

where

$$\sigma_{j1} = \sigma = I_j \tau \frac{1}{2} \left(\frac{\ell}{\sqrt{V_r}} \frac{1}{4\pi dc} \sqrt{\frac{\omega_k}{\hbar}} \right)$$

$$\sigma_{j0} = 0$$

$$\sigma_{ks} = 0; \qquad \text{for } j \neq k \text{ and } s = 0, 1$$

$$\theta_{ks} = \phi_{ks} - \omega_k \cdot \frac{d}{c} - \frac{1}{2}; \qquad \theta = \theta_{j1}$$

at all time t after the closure of the shutter. When the input message is m_s, the total received field is in the state given by the density operator ρ_s, where ρ_s in the P-representation is

$$\rho_s = \int P_s(\{\alpha_k\}) \; |\{\alpha_k\}\rangle\langle\{\alpha_k\}| \; \prod_k d^2\alpha_k$$

$$P_s(\alpha_k) = \frac{1}{\pi\langle n_j\rangle} \exp\left[-\frac{|\alpha_j - \sigma e^{-i\theta}|^2}{\langle n_j\rangle}\right] \prod_{k\neq j} \frac{1}{\pi \; n_k} \exp\left[-\frac{|\alpha_k|^2}{\langle n_k\rangle}\right].$$

Since modes other than the j^{th} contain no signal energy, and different modes are statistically independent, these modes can be discarded by the optimum receiver. It follows that the optimum receiver consists of only a resonant cavity with natural frequency, ω_j. The state of the field inside of this cavity after the shutter is closed is given either by ρ_0 or ρ_1, depending on whether m_0 or m_1 is transmitted. The corresponding weight functions in the P-representations of these density operators are

$$P_0(\alpha) = \frac{1}{\pi\langle n_j\rangle} \exp\left(-\frac{|\alpha|^2}{\langle n_j\rangle}\right)$$
$$P_1(\alpha) = \frac{1}{\pi\langle n_j\rangle} \exp\left(-\frac{|\alpha e^{-i\theta} - \sigma|^2}{\langle n_j\rangle}\right). \tag{9}$$

We suppose that the receiver measures the variable X of the received field inside the resonant cavity. On the basis of the outcome of the measurement, the receiver will make an estimation of the transmitted message, m. This variable associated with the j^{th} mode is represented by the linear operators, which are also denoted by X. The variable is not necessarily an observable as defined in quantum theory because we shall also allow variables that are represented by non-Hermitian operators. By the measurement of a complex variable corresponding to a non-Hermitian operator, we mean the "simultaneous measurement" of the two conjugate variables that are its real and imaginary parts. These measurements are not perfect ones in the conventional sense of quantum theory but result instead in two random variables, the product of whose variances satisfies the uncertainty principle.

We shall require that the operator X possess right eigenstates which form a complete set. That is, the equation $X|x\rangle = x|x\rangle$ has a nontrivial solution and $\int |x\rangle\langle x|\, dx = I$, where I is the identity operator. Such a class of operators includes all of the quantum-theoretical observables of a single mode. It also allows some "noisy simultaneous measurements" of conjugate variables which can be performed physically, for example, that of the amplitude and the phase of the field as done by a maser amplifier (where the operator X is then the annihilation operator a). The best receiver found from such a class of receivers may actually be one that measures variables of the field corresponding to the Hermitian operator as in the example discussed in the last part of this report.

The possible outcomes of the measurement of X are its eigenvalues $\underline{x}$. When X is non-Hermitian, its eigenvalue, $\underline{x}$, can also be regarded as two dimensional vector (Re x, Im x). Later on, whenever we speak of the probability density of $\underline{x}$, we mean either that of x if x is real or the joint probability density function of Re x and Im x if x is complex. Immediately after the completion of the measurement, the j^{th} mode of the field will be in the state $|x\rangle$, which is the right eigenstate of X associated with the eigenvalue $\underline{x}$. For simplicity, we shall assume that the eigenstates of the operator X are nondegenerate. If X is a Hermitian operator, the measurement process is just a quantum observation.

The outcome of the measurement of the received field is the random variable $\underline{x}$ defined over the sample space which contains all the eigenvalues of the operator X. Given that the input message is m_s, the conditional probability density function of the outcomes of the measurement of X (denoted as $p(x/m=m_s)$) is given by $\langle x| \int_s |x\rangle$; s = 1, 0. In terms of the weight function $P_s(\alpha)$, it is just

$$p(\underline{x}/m=m_s) = \int P_s(\alpha)\, |\langle \alpha|\, x\rangle|^2\, d^2\alpha. \tag{10}$$

It is convenient to expand the operator X which characterizes the receiver and its right eigenstate $|\underline{x}\rangle$ in terms of the coherent states

$$|x\rangle = \int \exp\left[-\tfrac{1}{2}|\alpha|^2\right] f_x(\alpha^*)\, |\alpha\rangle\, d^2\alpha$$

$$X = \frac{1}{\pi^2} \int \exp\left[-\tfrac{1}{2}|\alpha|^2 - \tfrac{1}{2}|\beta|^2\right] |\alpha\rangle X(\alpha^*, \beta)\, \langle\beta|\, d^2\alpha d^2\beta, \tag{11}$$

where

$$f_x(\alpha^*) = \exp\left[\frac{1}{2}|\alpha|^2\right] \langle\alpha|x\rangle$$

$$X(\alpha^*, \beta) = \langle\alpha|X|\beta\rangle \exp\left[\frac{1}{2}|\alpha|^2 + \frac{1}{2}|\beta|^2\right].$$

The equation $X|x\rangle = x|x\rangle$ can be rewritten in terms of $X(\alpha^*, \beta)$ and $f_x(\alpha^*)$ as the following integral equation:

$$f_x(\alpha^*) = \frac{1}{\pi x} \int X(\alpha^*, \beta)\, f_x(\beta^*) \exp[-|\beta|^2]\, d^2\beta. \tag{12}$$

In terms of this expansion for this binary signal, the conditional probability density functions of the observed result $\underline{x}$ when a measurement of X is made on the received field are given by the following equations

$$\begin{aligned} p(x/m_1) &= \frac{1}{\pi} \int P_1(\alpha)\, |f_x(\alpha^*)|^2\, d^2\alpha \\ p(x/m_0) &= \frac{1}{\pi} \int P_0(\alpha)\, |f_x(\alpha^*)|^2\, d^2\alpha. \end{aligned} \tag{13}$$

When $\underline{x}$ is observed, the maximum-likelihood receiver computes the conditional probability density functions $p(x/m_1)$ and $p(x/m_0)$, then sets $\hat{m} = m_1$ if $p(\underline{x}/m_1) > p(\underline{x}/m_0)$ (and sets $\hat{m} = m_0$ if otherwise). In terms of the weight functions, this condition is simply

$$\int (P_1(\alpha) - P_0(\alpha))\, |f_x(\alpha^*)|^2\, d^2\alpha > 0. \tag{14}$$

Let R_0 denote the region in the sample space of x in which the inequality (21) is not satisfied, and R_1 the region in which it is satisfied. The probability of detection error when the optimum receiver is used is

$$P(\epsilon) = \frac{1}{2}\left\{\int_{R_0} p(\underline{x}/m_1)\, d\underline{x} + \int_{R_1} p(\underline{x}/m_0)\, d\underline{x}\right\}. \tag{15}$$

Therefore, the problem of finding the optimum receiver is reduced to that of finding the optimum operator X (or, correspondingly, the optimal function $X(\alpha^*, \beta)$ that minimizes $P(\epsilon)$).

An interesting special case of this problem is the determination of the best receiver among a class of receivers corresponding to the operators, a, a^+, $X_{10}a^+ \pm X_{10}^*a$, or functions of one of these operators. This class of receivers has been extensively discussed because some physical devices correspond to one of these operators. It is most easily treated by expressing the weight functions of Eq. 9 as

$$P_1(\alpha_i, \alpha_q) = \frac{1}{\pi \langle n \rangle} \exp\left(-\frac{1}{\langle n \rangle}\left[\alpha_q^2 + \{\alpha_i - \sigma\}^2\right]\right)$$
$$P_0(\alpha_i, \alpha_q) = \frac{1}{\pi \langle n \rangle} \exp\left(-\frac{1}{\langle n \rangle}\left[\alpha_q^2 + \alpha_i^2\right]\right). \qquad (16)$$

α_i, in these equations, denotes the amplitude of the electric field in phase with the transmitted signal $|\alpha| \cos(\phi-\theta)$, as shown in Fig. XX-3. α_q denotes the quadrature phase

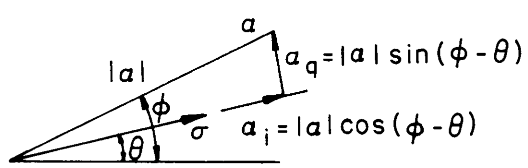

Fig. XX-3. Illustrating α_i and α_q.

term $|\alpha| \sin(\phi-\theta)$ of the electric field. The measurements of α_i and α_q correspond to that of the operators $X_{10}a^+ + X_{10}^*a$ and $X_{10}a^+ - X_{10}^*a$, respectively. The measurement in which α_i and α_q are measured simultaneously in such a manner that the product of the variances of α_i and α_q satisfy the uncertainty principle with equality corresponds to the operator a (or a^+).

It can be shown that when α_i and α_q are thus measured simultaneously, the outcome of the measurement of α_i will be statistically independent of α_q. Since α_q contains no signal energy, it can be discarded by the optimum receiver, according to the irrelevance theorem. Furthermore, it can be shown that a receiver that measures only the in-phase term yields a smaller error probability than the receiver that makes the simultaneous measurements and discards the quadrature phase term. Thus the optimum receiver in the class considered here should measure the in-phase component of the complex amplitude of the electric field in this mode, that is, the operator $\frac{1}{2}[e^{+i0}\, a + a^+\, e^{-i}]$. (It appears that the receiver is also better compared with the one that makes measurements of the energy of the received field. But the rigorous proof needs to be completed.) Such a measurement can be achieved by using a degenerate parametric amplifier.

If we regard the weight function $P(\alpha)$ as a probability density function, then the receiver that measures the in-phase component of the electric field is the maximum-likelihood receiver. Unfortunately, since the projection operators $|\alpha\rangle\langle\alpha|$ are not orthogonal to one another for different values of α, it is not always possible to interpret $P(\alpha)$ as a probability density function. If, however, we are given the knowledge that the field is in a coherent state of unknown α, the function $P(\alpha)$ may be thought of as playing a role analogous to the probability density function of the values of α over the complex

α-plane. Also, in the classical limit, when the function $P(\alpha)$ tends to vary little over a range of α, it can also be interpreted approximately as a probability density function. The former is the case when no additive noise is present, while the latter is the case of the classical limit $\langle n\rangle \gg 1$. In both cases, this receiver is a maximum-likelihood receiver.

The probability density functions of the measured value of α_i conditioned on the two messages transmitted are

$$p(\alpha_i/m_1) = \frac{1}{\sqrt{\pi\langle n\rangle+\frac{1}{2}}} \exp\left[-\frac{-(\alpha_i-\sigma)^2}{\langle n\rangle+\frac{1}{2}}\right]$$
$$p(\alpha_i/m_0) = \frac{1}{\sqrt{\pi\langle n\rangle+\frac{1}{2}}} \exp\left[-\frac{\alpha_i^2}{\langle n\rangle+\frac{1}{2}}\right]. \tag{17}$$

These equations are exactly the same as that in the case in which a classical signal of amplitude equal to either σ or 0 is transmitted in the presence of additive zero-mean Gaussian noise, with variance $\langle n\rangle + \frac{1}{2}$. Hence, much of the results on the computation of error probability obtained classically $\Leftrightarrow$ can be applied here. In particular, a simple analysis will show that the probability of error is given by

$$P(\epsilon) = Q\left(\frac{\sigma}{\sqrt{2\langle n\rangle+\frac{1}{2}}}\right) = Q\left(\sqrt{\frac{E_r}{2(N_T+N_z)}}\right), \tag{18}$$

where

$E_r = \sigma^2\hbar\omega$ is the average received energy.

$N_T = \langle n\rangle\hbar\omega$ is the energy in the thermal noise.

$N_z = \frac{1}{2}\hbar\omega$ is the quantum noise energy, which is usually called by physicists the zero-field fluctuation energy.

$$Q(x) = \frac{1}{\sqrt{2\pi}}\int_x^\infty \exp\left[-\frac{1}{2}y^2\right] dy.$$

This probability of error equals that obtained by Helstrom[2] for a receiver that makes simultaneous measurements of α_i and α_q. He used the Weigner distribution of α_i and α_q, instead of their joint probability density function to compute the value of the likelihood ratio. In his more recent paper[3], he shows that, for weak signals, a threshold detector also measures α_i with probability of error given by Eq. 18 for coherent binary signals.

The performance of this receiver when used with other types of signals can also be found easily. For example

i. Binary antipodal signals: $P(\epsilon) = Q\left(\sqrt{\frac{2E_r}{(N_T+N_z)}}\right)$.

ii. Binary orthogonal signals: $P(\epsilon) = \int_{-\infty}^{\infty} dx \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(x - \sqrt{\frac{E_r}{2(N_T+N_z)}}\right)^2\right][Q(x)]$.

iii. M orthogonal signals: $P(\epsilon) = \left(1 - \int_{-\infty}^{\infty} dx \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(x - \sqrt{\frac{E_r}{2(N_T+N_z)}}\right)\right][1-Q(x)]^{M-1}\right)$.

The model of the receiving system discussed here will be generalized in two directions. First, since we are interested in the changes in the statistics of the measured values and in the accuracy of our estimations with respect to the length of time spent in observing the received field, the model of the receiver will be modified so that the effect of the observation time is included. Second, since it has not been proved that the optimum receiver, indeed, observes the individual modes of the received field separately, the model will be generalized so that multi-mode measurements are also allowed. Several new problems arise in the multi-mode detection of arbitrary coherent signals: the optimum design of the receiving cavity, the effect of the correlation between modes, the dependence of the performance of the optimum receiver upon the set of signals and the properties of an optimum signal set.

Jane W-S. Liu

References

1. R. J. Glauber, "Coherent and Incoherent States of the Radiation Fields," Phys. Rev. 131, 2766-2788((1963).

2. C. W. Helstrom, "Quantum Limitations on the Detection of Coherent and Incoherent Signals," IEEE Trans., Vol. IT-11, No. 4, pp. 482-490, 1965.

3. C. W. Helstrom, "Detection Theory and Quantum Mechanics," Inform. Contr. 10, 254-291 (1967).

C. BLOCK CODES FOR DISCRETE SOURCES AND CERTAIN FIDELITY CRITERIA

The encoding by means of block codes of discrete memoryless information sources to satisfy a fidelity criterion has been investigated for the case in which the distortion between a source letter x and a letter y in the alphabet to be used to represent the source is either zero or infinite. This report derives an expression for the smallest number of code words such a block code may have when the fidelity criterion is that the average distortion between source outputs and reconstructed letters be zero. Knowledge of this minimum block code rate for such a distortion measure allows us to solve several related problems involving block codes. For example, we can specify the range of rates for which such codes can approach the Rate-Distortion function in performance, and we can calculate the rate of a discrete source relative to a fidelity criterion that requires every letter to be encoded with distortion no more than a specified amount, rather than merely achieving this performance on the average. The criterion on average distortion with letter distortions either zero or infinite is equivalent to requiring that for every possible source sequence, there be a code word whose distortion with this sequence is zero. The distortion between blocks is taken to be the sum of the individual letter distortions.

Thus for each source letter x, there is a set of "allowable" output letters, and one of these must be used to adequately represent x. These sets can be represented simply by an adjacency diagram like that shown in Fig. XX-4, which is interpreted to mean that any output letter that is connected by a line to an input letter may be used to represent that input letter. In both of these examples, the input and output are 3-letter alphabets, but in A, only the corresponding letter may be used to represent a source output, while in B, the first source letter may be coded into (represented by) either of the first two putput letters, etc. For the first case, it is easy to see that we need one code word for each possible source sequence, and so will need 3^n words to insure that blocks of n will be adequately represented. In the second case, however, the set $\{1,2\}$ is a satisfactory set of code words for n = 1, and $\{11,22,33\}$ is such a set for n = 2. We shall see that the rate of this second code, $R = \frac{1}{2} \log 3$, is the smallest that any satisfactory code can have.

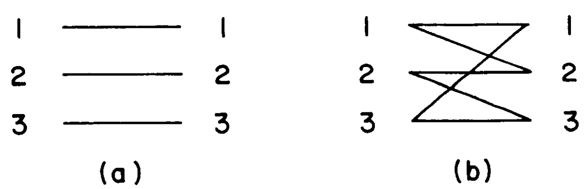

Fig. XX-4. Adjacency diagram.

Let us now turn to the formal development of our results, and for a block length of n, define $M_{o,n}$ as the smallest number of code words that insures that every source sequence can be represented by at least one code word. Then we define the rate, R_o, by

$$R_o \triangleq \inf_n \frac{1}{n} \log M_{o,n}.$$

Clearly, $R_o \geq R(0)$, the rate-distortion function for the given source and distortion measure evaluated at D = 0. This rate-distortion function is defined by

$$R(D) = \min_{p(j|i)} I(X;Y),$$

where the minimization is done subject to the constraint that the average distortion be at most D, that is,

$$\sum_{ij} p_i p(j|i)\, d_{ij} \leq D,$$

where d_{ij} is the distortion between source letter i and output letter j, and, in our case, is either zero or infinity. Shannon[1] has shown that no coding scheme that gives an average distortion of D can have a rate below R(D), and conversely, when all distortions are finite there are schemes that have rates arbitrarily close to R(D). It has further been found that even when some distortions are infinite, there are coding schemes whose average distortions and rates approach D and R(D), respectively, but these schemes may not be block ones.[2]

Now, if we require all of the source letter probabilities to be nonzero, which merely states that all letters are really there, then it is easy to see that $M_{o,n}$ is independent of these probabilities. This is true, since the probability of each source sequence will then be nonzero, and so must have a representative code word. Thus the inequality $R_o \geq R(0)$ must hold for all source distributions, and so we can write

$$R_o \geq \sup_{\underline{p}} R(0)$$

where $\underline{p}$ is the source probability vector, and the sup is over the open set described by the conditions $p_i > 0$ for all i, and $\sum_i p_i = 1$. Now, since I(x;y) is a continuous function of the probability distributions, so is R(D), and we can therefore include the boundary of the region and write

$$R_o \geq \max_{\underline{p}} R(0),$$

where now the max is over the set $p_i \geq 0$, and $\sum_i p_i = 1$. Finally, inserting the defining relation for R(0), we have

$$R_o \geq \max_{\underline{p}} \min_{p(j|i)} I(x;y), \tag{1}$$

where the transition probabilites $p(j|i)$ of the test channel are restricted to be nonzero only for those transitions that result in zero distortion (i. e., values of i and j for which the distortion between i and j is zero).

We shall now show that the inequality of Eq. 1 is actually an equality, by demonstrating the existence of satisfactory codes with rates equal to the right-hand side. Since the extrema of functions over closed sets are always attained for some member of the set, there must exist $\underline{p}^*$ and $p^*(j|i)$ such that the mutual information that they determine is the desired max min. Since I(X;Y) is a differentiable function of the probability distributions, it must be stationary with respect to variations in the $p(j|i)$ distribution at the saddle point. If this were not so, then this point could not be a minimum. Furthermore, since I(X;Y) is convex $\cap$ in $\underline{p}$ and convex $\cup$ in the $p(j|i)$ distribution, the order of the max and min can interchanged without affecting the result.[3] Then, by the same reasoning as that above, I(X;Y) must also be stationary with respect to variations in the nonzero p_i when $p_i = p_i^*$.

Thus including Lagrangian multipliers to satisfy the constraints that probability distributions sum to one, we can write

$$\frac{\partial}{\partial p(j|i)}\left[I(X;Y) + \sum_{ij} \lambda_i p(j|i) + \mu \sum_i p_i\right] = 0 \tag{2}$$

and

$$\frac{\partial}{\partial p_i}\left[I(X;Y) + \sum_{ij} \lambda_i p(j|i) + \mu \sum_i p_i\right] = 0 \tag{3}$$

for all i and j such that p_i^* and $p^*(j|i)$ are nonzero. Since we know

$$\frac{\partial I}{\partial p(j|i)} = p_i \log \frac{p(j|i)}{q_j}$$

and

$$\frac{\partial I}{\partial p_i} = \sum_j p(j|i) \log \frac{p(j|i)}{q_j} - 1,$$

where $q_j = \sum_i p_i p(j|i)$, then Eq. 2 becomes

$$p_i \log \frac{p(j|i)}{q_j} + \lambda_i = 0$$

or

$$\log \frac{p(j|i)}{q_j} = r_i,$$

which is independent of j. Then by substituting r_i for $\log \frac{p(j|i)}{q_j}$, Eq. 3 becomes

$$\sum_j p(j|i)\, r_i - 1 + \mu = 0,$$

and by taking r_i outside the summation,

$$r_i = 1 - \mu$$

and thus is independent of i also. Therefore for the maximizing and minimizing distributions, it must be true that

$$\frac{p(j|i)}{q_j} = h^{-1}$$

a constant, for all i and j such that $p_i p(j|i) > 0$. It then follows that

$$I(X;Y) = \log \frac{1}{h},$$

and if we define the sets $S_i = \{j | d_{ij} = 0\}$, we can write

$$\sum_{j\in S_i} q_j = \sum_{j \in S_i} hp(j|i) = h \qquad (\text{all } i \text{ such that } p_i > 0)$$

since $p(j|i)$ must be zero for $j \notin S_i$.

For those source letters whose probabilities turn out to be zero, an expression similar to Eq. 3 must hold, namely

$$\frac{\partial}{\partial p_i}\left[I(X;Y) + \sum_{ij} \lambda_i p(j|i) + \mu \sum_i p_i\right] \leq 0 \tag{4}$$

for all such that $p_i = 0$. By taking the derivative, this becomes

$$\sum_j p(j|i) \log \frac{p(j|i)}{q_j} - 1 + \mu \leq 0$$

or

$$\sum_j p(j|i) \log \frac{p(j|i)}{q_j} \leq 1 - \mu = \log \frac{1}{h}. \tag{5}$$

Since the $p(j|i)$ do not affect $I(X;Y)$ when p_i is zero, this inequality must be true for all choices of the $p(j|i)$ values, as long as these are zero for $j \notin S_i$. So in particular, it must hold when

$$p(j|i) = \begin{cases} \frac{q_j}{\sigma}, & j \in S_i \\ 0 & \text{otherwise} \end{cases}$$

where σ is a normalizing constant $= \sum_{j \in S_i} q_j$. But with this choice of $p(j|i)$, Eq. 5 becomes

$$\log \frac{1}{\sigma} \leq \log \frac{1}{h}$$

or

$$\sigma = \sum_{j \in S_i} q_j \geq h.$$

Now consider constructing a code by picking each letter of each code word randomly and independently, with probability distribution q_j. With this method, the probability that a randomly selected letter will be an acceptable representation of the i^{th} source letter is just $\sum_{j \in S_i} q_j$, which has been shown to be greater than or equal to h, independent of i.

Thus the probability that a randomly chosen code word y of block length n will be acceptable for a given source sequence $\underline{x}$ is

$$\Pr [\underline{y} \sim \underline{x} | \underline{x}] \geq h^n$$

for all $\underline{x}$.

If $M = e^{nR}$ code words are so chosen, then the code rate is R, and the probability (averaged over all codes) that a given $\underline{x}$ sequence is not covered is

$$\Pr [\text{no word for } \underline{x} | \underline{x}] \leq (1-h^n)^M$$

$$\leq e^{-Mh^n} = \exp\left[-e^{n\left(R - \log \frac{1}{h}\right)}\right].$$

Now setting $R = \log \frac{1}{h} + \delta$, where δ is an arbitrarily small positive number, we can bound the probability of all $\underline{x}$ sequences for which there is no code word by

$$\Pr[\underline{x} \mid \exists \text{ no code word for } \underline{x}] = \sum_{\underline{x}} p(\underline{x}) \Pr[\text{no word for } \underline{x} \mid \underline{x}] \leq e^{-e^{n\delta}}.$$

By the usual random coding argument, there must exist a code with performance at least as good as this average, and we can choose n large enough so that

$$\frac{e^{-e^{n\delta}}}{p_{min}^{n}} < 1,$$

where p_{min} is the smallest source letter probability. Thus for large enough n, there is a code for which

$$\Pr[\underline{x} \mid \exists \text{ no code word for } \underline{x}] \leq e^{-e^{n\delta}} < (p_{min})^{n}.$$

But this probability must then be zero, since the smallest nonzero value it could take on is p_{min}^{n}.

Thus we have shown that for code rates arbitrarily close to $\log \frac{1}{h} = \max \min I(X;Y)$, there exist codes that will give zero distortion when used with the given source. Since it has already been shown that such performance was not possible for rates below this value, we have established the block-coding rate of a discrete source with letter distortions that are either zero or infinite to be

$$R_o = \max_{\underline{p}} \min_{p(j|i)} I(X;Y).$$

Furthermore, the conditions on stationarity of I given by Eqs. 2, 3, and 4, are sufficient as well as necessary. Thus we know that if we can find distributions satisfying the condition that for some number h, $\frac{p(j|i)}{q_j} = h^{-1}$ for all i and j such that $p_i p(j|i) > 0$, then the I(X;Y) determined by these distributions is the desired R_o. We have seen that this condition is equivalent to having all transitions leading to a given output letter have the same probability, and that for each source letter with positive probability, the sum of the probabilities of those output letters that can be reached from this input be the same.

It is interesting to compare the results above with those of Shannon on the zero error capacity of a noisy channel.[4] His upper bound on this capacity is identical with our R_o.

As an example of the calculation of R_o, consider the distortion shown in Fig. XX-4b.

It is easy to see that a source whose letter probabilities are all 1/3, and a channel whose transition probabilities are 1/2 everywhere there is a line in the diagram, and zero elsewhere, satisfy the conditions on stationarity given above. Thus R_o is I(X;Y) for this source and channel, which comes out to be $\frac{1}{2}\log 3$.

The techniques developed here allow us to solve several other problems concerning block codes for discrete sources. If we want to know the rate needed for zero average distortion for an arbitrary distortion matrix, it is easy to see that this rate is just R_o for a distortion that is zero everywhere the given one is, and infinite elsewhere. A second problem is finding the block-code rate necessary for achieving a finite average distortion when some of the d_{ij} are infinite. In this case, the answer is R_o for the distortion that is infinite everywhere the given one is, and is zero elsewhere. Finally, the fidelity criterion might be that every letter be encoded with a distortion no greater than D (as opposed to the more usual constraint on the average over a block, which we have been considering up to this point). In this case, the "every letter" block-coding rate $R_{el}(D)$ is just R_o for the distortion that is zero if the original d_{ij} is D or less, and infinite otherwise.

J. T. Pinkston III

References

1. C. E. Shannon, "Coding Theorems for a Discrete Source with a Fidelity Criterion," IRE National Convention Record, Part 4, pp. 142-163, 1959.

2. J. T. Pinkston, Doctoral thesis research (in preparation).

3. S. Karlin, Mathematical Methods and Theory in Games, Programming and Economics (Addison-Wesley Publishing Company, Reading, Mass. 1959), pp. 28-30.

4. C. E. Shannon, "The Zero Error Capacity of a Noisy Channel," IRE Trans., Vol. IT-2, No. 3, pp. 8-19, September 1956.

D. INFORMATION TRANSMISSION VIA SPATIAL MODULATION OF WAVEFORMS

In many communication systems the physical antenna system influences the system reliability only through the received signal-to-noise ratio (SNR).[1,2] In contrast, we focus on the consequences of modulating source outputs onto temporal and spatial variations of the "complex amplitude" of the signal illumination across the transmitter antenna aperture.

We shall apply information-theoretic concepts to determine ultimate limits on the potential for information transmission between optical devices. In this context we shall evaluate the "channel capacity" and "rate-reliability" function that is obtained when a transmitting antenna of given physical dimensions is driven with the best possible spatial and temporal modulation. We expect that this optimum modulation will depend upon the source information rate, physical dimensions of the antenna system, available signal power, and the characteristics of associated noise sources. Channel capacity and the optimum reliability will be computed and used as a standard for comparison with certain suboptimal modulation techniques.

Although the word "optics" carries the particular connotation of visible light frequencies, the problem that we pose is essentially independent of the radiation wavelength (for instance, we could talk about illuminating large array antennas at microwave frequencies). We expect, however, that encoding information in the spatial variation of EM fields is theoretically attractive only when the aperture diameters are several orders of magnitude larger than the carrier wavelength. This condition is not usually satisfied at microwave frequencies. Therefore we are, in fact, especially concerned with optical frequencies, and have in mind the potential application of laser technology to communication.

We treat that part of the communication link illustrated in Fig. XX-5. The channel includes a pair of physical apertures I and II, which we designate as transmitter and receiver. These apertures are coupled via radiation through free space. We presume that the transmitting aperture can be excited by any one of an arbitrary set of complex illuminations, say, $S_i(x, y, z, t)$. The input illumination is observed as the amplitude distribution, $O_i(a, b, c, t)$, at the receiving aperture. If follows from Maxwell's equations and the free-space assumption that O_i is a <u>linear</u> functional of S_i. To complete the description of Fig. XX-5, we have shown an additive noise, N, that represents the sum of all background radiation, incident on the receiving aperture, which is uncorrelated with the source illumination. Maxwell's equations ensure linearity for a wave propagating

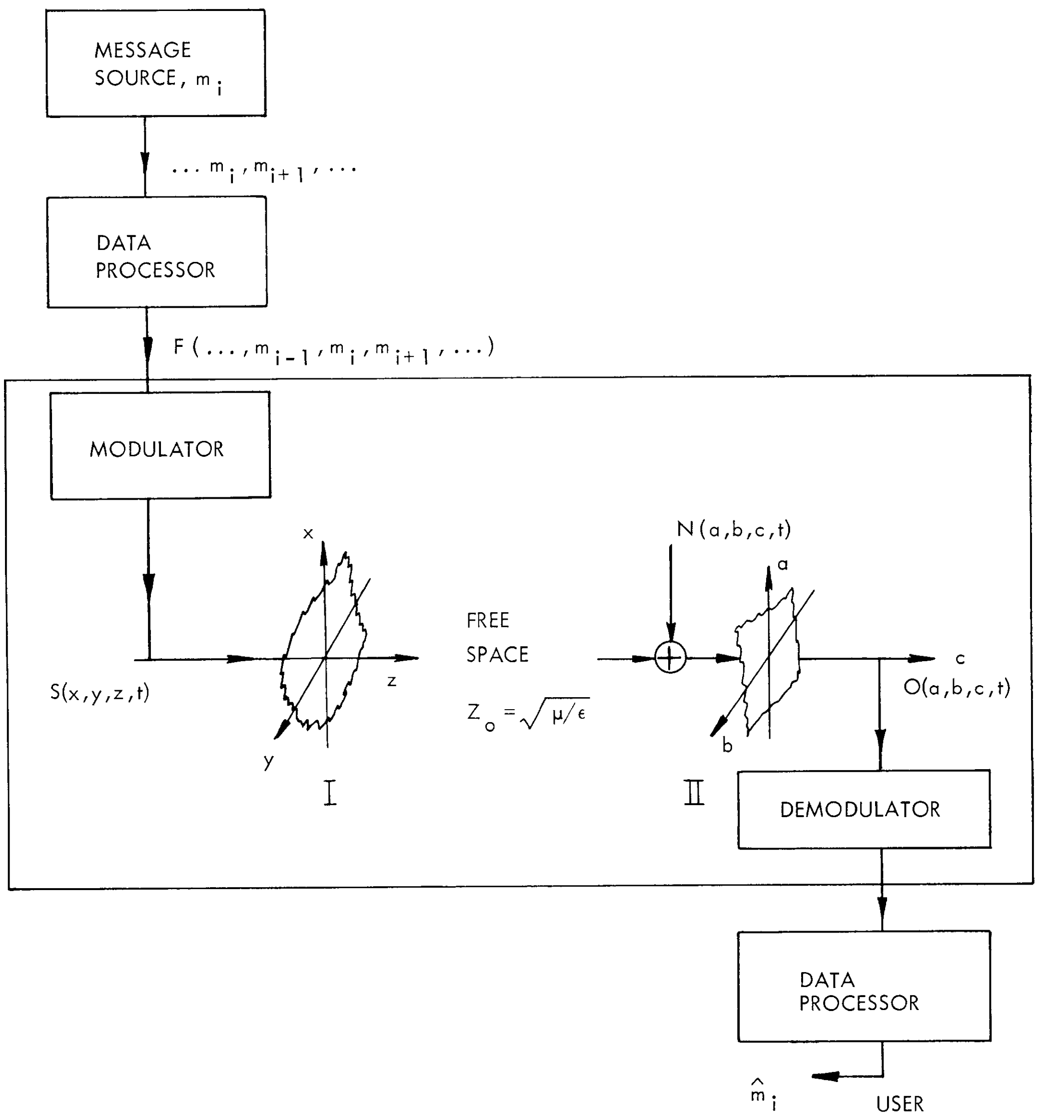

Fig. XX-5. The optical channel.

in free space. Nonlinearity may become an issue in certain optical systems, e.g., those that include photo-emulsions.

1. Antenna Input-Output Relations

The channel model is based on the representation of EM waves by solutions to Maxwell's equations. Analysis of radiation problems is usually reduced to the solution of an integral equation with associated boundary conditions imposed by sources, dielectric boundaries or conducting surfaces.[3-6] In general, one cannot solve such problems unless the geometry is particularly simple or certain simplifying assumptions are made. We can often go far by using simple solutions because Maxwell's equations are linear; hence, the principle of superposition applies.

The first step is to determine the (complex) amplitude distribution of the field O(·) across an image aperture, given the distribution of the field S(·) across an object (source) plane. For example, we typically specify that S(·) is nonzero only over some finite region of the infinite plane. Let this region be the surface Σ. Following Sommerfeld's "Manifestly Consistent" approach, it can be shown that $\underline{U}(P_0)$, which is the complex amplitude of the field at P_0 when the source aperture Σ is illuminated by an arbitrary excitation $\underline{U}(P_1)$, is given by

$$\underline{U}(P_0) = \iint_{\Sigma} \underline{U}(P_1) \frac{e^{j\beta r_{01}}}{j\lambda r_{01}} \cos(\vec{n}, \vec{r}_{01}) \, ds, \tag{1}$$

where

$\vec{n}$ is an outward normal from Σ

$\vec{r}_{01}$ is the vector from P_1 to P_0

$\beta = \frac{2\pi}{\lambda}$

λ = wavelength of the monochromatic light.

We can write Eq. 1 in the alternative form of a linear system response.

$$\underline{U}(P_0) = \iint_{\Sigma} h(P_0, P_1) \, \underline{U}(P_1) \, ds, \tag{2}$$

where

$$h(P_0, P_1) = \frac{1}{j\lambda} \frac{e^{j\beta r_{01}}}{r_{01}} \cos(\vec{n}, \vec{r}_{01}).$$

In order to carry this analysis further, we have to make some simplifying

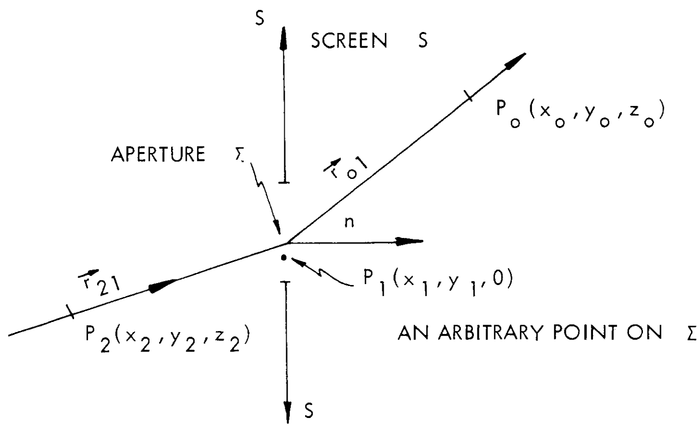

Fig. XX-6. Aperture geometry for diffraction analysis.

approximations to the impulse response $h(P_0, P_1)$. These are usually attributed to Fresnel. Using the notation of Fig. XX-6, we have

$$z \gg \left(x_1^2+y_1^2\right)^{1/2} \max$$

$$\cos(\vec{n}, \vec{r}_{01}) \approx 1.$$

Substituting these expressions in Eq. 1, we find for all points P_0 that are "close to the optical axis" (that is, within approximately 18°)

$$r_{01} = \left(z^2+(x_0-x_1)^2+(y_0-y_1)^2\right)^{1/2}$$

$$\approx z\left(1 + \frac{(x_0-x_1)^2}{2z^2} + \frac{(y_0-y_1)^2}{2z^2}\right)$$

$$\underline{U}(P_0) = \frac{e^{j\beta z}}{j\lambda z} \iint_{\Sigma} \underline{U}_1(X, Y) \exp\left(\frac{j\beta}{2z}\left\{(x_0-x_1)^2 + (y_0-y_1)^2\right\}\right) dxdy \tag{3}$$

Thus, the received field can be modeled by the output of a linear, spatially invariant filter,[7] $h_\lambda(P_0, P_1)$, where

$$h_\lambda (P_0, P_1) = \frac{\exp j\beta z}{j\lambda z} \exp \frac{j\beta}{2z} \left|P_0 - P_1\right|^2. \tag{4}$$

2. Reliability of the Optical Communication Channel

We note that Maxwell's equations imply that the radiation field is a linear functional of the illumination at the source aperture. This linearity can be exploited to apply the results of Information Theory to an analysis of ultimate limits on the communication performance of radiating systems.

Several authors, following Shannon,[8-10] have shown that for each pair of channel impulse response and additive (Gaussian) noise power spectrum there is an optimum set of average-power limited message waveforms. Their criterion of performance is the average probability of receiving a message in error. These studies have shown that this probability may be bounded by expressions of the form

$$A(R) \exp(-NE_L(R)) \leq P_e \leq \exp(-NE_U(R)),$$

where R is the information rate per channel use relative to channel capacity, and N is the number of channel uses per message.

We are especially interested in extensions of this work published recently by Holsinger[9] and Ebert.[10] Their work related to the analysis of signalling through time-continuous and amplitude-continuous channels with memory, that is, the filter channel.

Holsinger derives a complete set of orthogonal, finite duration waveforms which serve as a coordinate system in which one can expand the inputs to a filter channel. These waveforms have the special property that they remain orthogonal at the output of the channel, although their functional dependence on time has been altered. (A general expansion theory applied to optical transmission has been presented by Gamo.)

This orthogonal-function expansion suggests that the filter channel be represented by the parallel combination of independent channels each with a different mean signal strength. Holsinger and Ebert studied the problem of optimal coding for such parallel channel combinations and their results include the tightest available exponential bounds to the rate-reliability curves for the average-power limited transmitter.

Our interest lies in applying these results to optical systems. That they apply follows from the linearity of the optical impulse response as seen from Eq. 3. Therefore we can determine a set of two-dimensional waveforms in which to expand the spatial dependence of radiated signals such that we can isolate spatially orthogonal waves at the transmitter and receiver. We may then apply Ebert's results to obtain performance bounds for these parallel channels.

We are less concerned with the formal mathematics (except for making necessary modifications) than with the way the computed reliability functions vary with the physical properties of the channel such as aperture sizes, optical processing, and noise characteristics.

3. The 1-dimensional Antenna

We can modify Eq. 3 to obtain the signal component of the 1-dimensional antenna output. (Signals are represented by their complex envelopes. Thus the time variation of u(x, t) has been suppressed, and we indicate explicitly only the spatial modulation u(x).)

$$\underline{U}(Y) = \frac{e^{j\beta z}}{j\lambda z} \int_{-D_1}^{D_1} \underline{U}_1(X) \exp \frac{j2\pi(x-y)^2}{2\lambda z}\, dx.$$

The phase term is independent of y; we assume that it can be compensated for in the receiver processing. This compensation will not affect the relevant statistics of the additive receiver noise.

It can be shown that the appropriate expansion functions are $\{\psi_i(y)\}$ and $\{g_j(x)\}$, where

$$\psi_i(y) = \phi_i(y)\, e^{\frac{j\beta}{z} y^2}$$

$$g_j(x) = \frac{\beta}{z} \int_{-\infty}^{\infty} \psi_j(y)\, e^{\frac{-j\beta}{2z} (y-x)^2} \frac{dy}{2\pi}$$

and $\{\phi_i(y)\}$ are solutions to the following integral equation

$$\int_{-D_2}^{D_2} \frac{\sin \frac{\beta}{z} D_1 (x-y)}{\pi(x-y)} \phi_i(x)\, dx = K_i \phi_i(y) \qquad \text{for all } y \tag{5}$$

$2D_1$ = length of input aperture

$2D_2$ = length of output aperture

We recognize the $\{\phi_i\}$ to be the prolate-spheroidal wave functions discussed by Pollak and Slepian,[12, 13] among others, who have listed the following useful properties:

1a- $$\int_{-\infty}^{\infty} \phi_i(t)\, \phi_j(t)\, dt = 0 \qquad i \neq j$$
$$= 1 \qquad i = j$$

1b- The ϕ_i are bandlimited. Thus $\{f_j(x)\}$ are complete on the interval $|x| \leq D_1$

$$f(x) = \int_{-\infty}^{\infty} \frac{\beta}{z} \phi(y) e^{\frac{j\beta}{z} xy} \frac{dy}{2\pi}.$$

2- The $\{\phi_i\}$ are complete on $|y| \leq D_2$.

It follows that

$$g(x) = e^{\frac{-j\beta x^2}{2z}} f(x).$$

Therefore properties 1 and 2 are also satisfied by the $\{\psi_i\}$ and $\{g_i\}$. It follows from Parseval's theorem and the relations above, that the $\{g_i\}$ is a complete set of functions that we can scale to have unit energy on the interval $|x| \leq D_1$.

We cannot proceed further without specific information about the eigenvalues $\{K_i\}$. The key parameter that controls these eigenvalues is N, where

$$N = \frac{\beta}{z} D_1 D_2 = \frac{2\pi}{\lambda z} D_1 D_2 \tag{6}$$

Pollak and Slepian show that the eigenvalues approach zero rapidly for $i \geq 2N/\pi$. When N gets large, say, $N \geq 50$, the distribution of eigenvalues tends to the sharp cutoff illustrated in Fig. XX-7.

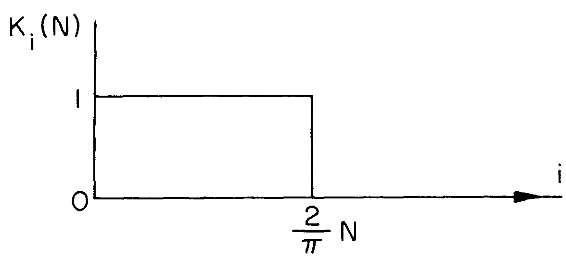

Fig. XX-7. Limiting distribution of eigenvalues.

In qualitative terms $\frac{2N}{\pi} = N'$ is the number of significant eigenvalues, and can be identified as the number of degrees of freedom for the system. When the channel output SNR is low, the available transmitter power should be used to drive the aperture with only the most significant eigenfunctions. Higher order eigenfunctions become important only as the SNR increases. In the limit of zero additive noise, each of the infinite set of eigenfunctions can be used. In other words, the concept of "degrees of freedom" is useful only when there is additive noise in the system. (Tufts[14] has remarked on this idea and gives a simple example in which more than 2W samples per second can be exactly recovered at the output of a noise-free lowpass channel of bandwidth W.)

Equation 6 gives an explicit relation for N in terms of the properties of the antenna system. For fixed aperture sizes, N decreases with the distance z. When the receiver is in the far field of the transmitter there is only one dominant eigenvalue, unless the receiving aperture is much larger than the transmitting aperture. This follows directly from the far-field condition

$$z \gg \frac{\pi D_1^2}{\lambda}.$$

When there is only one significant eigenvalue, all available signal power should be used to excite the corresponding eigenfunction, unless the SNR is "high." Even then, the increase in channel capacity, with the "weaker" coordinates used, will be small.

In typical microwave applications antennas are sited in the far field and are not large enough for N' to exceed 1. It follows that the potential gain from using spatial modulation at microwave frequencies is negligible. Of course, there may be some benefit from optimum selection of the single illumination pattern (optimum beam forming) in terms of maximizing the energy incident on the receiving aperture.

4. Interpretation and Future Work

The representation of the radiating part of the optical communication link by the parallel combination of independent channels leads directly to a solution for the set of optimum input signals. The transmitted waveforms are constructed as linear combinations of temporally modulated spatial eigenfunctions. The temporal modulation, in turn, can be represented as a linear combination of eigenfunctions, as described by Landau, Pollak, and Slepian. The optimum distribution of average power amongst these coordinates is obtained from a "water-pouring" argument based on the relative amplitudes of the corresponding eigenvalues.

The analysis outlines here treats the case of optimum signaling and coherent processing of EM fields across the receiver aperture. In view of practical difficulties of implementing optical coherent detectors, we wish to estimate the performance degradation that one accepts as the price of using suboptimum processing. In particular, we have in mind using square-law (energy) detection, which is implemented at optical frequencies via photographic emulsion. We have recently begun this aspect of our study, and will consider various suboptimal techniques at the transmitter and receiver.

Other work is in progress to establish the interplay between spatial modulation and the maximum temporal signaling rate. This work focuses on estimating the rate at which one can establish in succession arbitrary radiation patterns, subject to maintaining a high ratio of radiated energy to energy stored in the EM fields around the source aperture.

We may consider two directions in which to extend these studies. The first relates to taking into account the quantum noise introduced by certain detectors. We believe this noise may be the principal additive disturbance, at low-rate signaling. The second is to introduce atmospheric turbulence into our channel model. At present, we expect that this would be a major extension, beyond the scope of the present problem.

R. L. Greenspan

References

1. C. E. Shannon, "A Mathematical Theory of Communication," Bell System Tech. J. 27, 379 (1948).
2. R. M. Fano, Transmission of Information (The M. I. T. Press, Cambridge, Mass., 1961).
3. J. A. Stratton, Electromagnetic Theory (McGraw-Hill Book Company, New York, 1941).
4. R. B. Adler, L. J. Chu, and R. M. Fano, Electromagnetic Energy Transmission and Radiation (John Wiley and Sons, Inc., New York, 1960).
5. M. Born and E. Wolf, Principles of Optics (Pergamon Press, New York, 1964).
6. R. Luneberg, Mathematical Theory of Optics (University of California Press, Berkeley, California, 1964).
7. J. Winthrop and C. Worthington, "Convolution Formulation of Fresnel Diffraction," J. Opt. Soc. Am. 56, 588-591 (1966).
8. R. M. Gallager, "A Simplified Derivation of the Coding Theorem," IEEE Trans., Vol. IT-11, No. 1, pp. 1-17, January 1965.
9. J. L. Holsinger, "Digital Communication over Fixed Time-Continuous Channels," Technical Report 430, Research Laboratory of Electronics, M. I. T., Cambridge, Mass., October 20, 1964.
10. P. M. Ebert, "Error Bounds for Parallel Communication Channels," Technical Report 448, Research Laboratory of Electronics, M. I. T., Cambridge, Mass., August 1, 1966.
11. H. Gamo, "Matrix Treatment of Partial Coherence," in Progress in Optics III, edited by E. Wolf (North-Holland Publishing Co., Amsterdam, 1964), p. 186 ff.
12. D. Slepian and H. Pollak, "Prolate Spheroidal Wavefunctions I.," Bell System Tech. J. 40, 43-64 (1961).
13. H. Landau and H. Pollak, "Prolate Spheroidal Wavefunctions III.," Bell System Tech. J. 41, 1295-1336 (1962).
14. D. M. Tufts, "Nyquist's Problem," Proc. IEEE 53, 248-260 (1965).