DIGITAL FM - TUTORIAL[1]

J. SALZ
Electrical Engineering Department
University of Florida.

Summary A review of the state of knowledge of digital FM techniques is undertaken. The digital FM signal and its spectral properties are first discussed. We then turn to the analysis of discrimination detection and review a recently proposed phenomenological model from which the error causing mechanism can be understood. We use this model to derive estimates of error-rate as a function of pertinent system parameters. The results obtained for practically instrumented systems are then compared with the ideal. The paper concludes with a discussion of some computer-aided analysis capable of predicting the performance of digital FM systems operating over the dispersive gaussian channel.

Introduction Frequency modulation (FM) techniques are widely used in telemetering systems and other data communications systems. Noncoherent digital FM techniques have particularly great appeal because of instrumentation advantages over other systems. The instrumentation advantages derive principally due to the noncoherent nature of the modulation process and the relative ease of signal generation. The various demodulators operate directly on the received signal-plus-noise and detection is achieved without the need to recover carrier phase as is usually done in linear modulation systems such as single or vestigial-sideband AM. However if optimum efficiency in data rate per unit bandwidth is desired, FM clearly is not the best technique. This is because FM is basically a double-sideband process requiring both upper and lower sidebands for efficient detection. We remark however that single-sideband FM has been investigated[(1),(2)] in the literature but the 3 dB saving in bandwidth promised theoretically cannot generally be realized in practice.

When bandwidth utilization is not the major consideration, FM does provide excellent performance with minimum equipment complexity. FM techniques are also immune, to some degree, to certain transmission medium impairments such as frequency and phase shifts and some types of nonlinearities. Moreover, since only the frequency of a sinusoid is modulated, the amplitude and consequently the power level of the modulated signal is constant. These attributes make FM well suited in applications where the peak power must be limited and independent of-the modulating signal.

---

[1] This work was supported by the National Aeronautics and Space Administration under Contract No. NsG-542 Supplement No. 4

In this paper we undertake to review the state of knowledge concerning performance of FM systems. We begin by discussing the digital FM signal and its spectral properties. We then turn to the analysis of discrimination detection and review a recently proposed phenomenological model from which the error causing mechanism can readily be understood. We use this model to derive error-rate vs signal-to-noise ratio formulas when signaling takes place over the additive gaussian channel. We then compare the attainable performance of practically instrumented systems with the best possible theoretical performance. We conclude the paper by reviewing some computer-aided analysis capable of predicting the performance of digital FM systems operating over the dispersive gaussian channel.

The Digital FM Signal and Spectral Properties A continuous-phase constant-envelope FM signal is customarily generated by varying the quiescent frequency of an oscillator in accordance with an information-carrying signal. These oscillators are commonly referred to as voltage-controlled oscillators (VCO). If we let x(t) represent the baseband information-carrying signal and ω_o the quiescent radian frequency of the oscillator, the output signal from the VCO may ideally be represented by

$$y(t) = A \cos \left[\omega_o t + \omega_d \int_0^t x(\tau)\, d\tau + \phi\right] \tag{1}$$

where A is the amplitude of the oscillator and ϕ is an initial phase angle. The constant ω_d is a conversion factor relating radians to units of x(t).

In this paper we shall be primarily concerned with digital FM waves and therefore we assume for the purposes of this discussion that x(t) can be represented by

$$x(t) = \sum_n a_n\, g(t - nT) \tag{2}$$

where (a_n) is a sequence of integers picked at random. The pulse g(t) satisfies

$$g(t) = \begin{cases} 1, & 0 \leq t \leq T \\ 0, & \text{elsewhere} \end{cases} \tag{3}$$

where 1/T is the baud (signaling rate). When the representation of the baseband signal as indicated in Eq. 2 is used in Eq. 1 we see that the wave y(t) is a frequency-shift-keyed (FSK) signal. The frequency in each T-interval is $\omega_d a_n + \omega_c$. For example if $a_n = \pm 1$, for all n, we have a binary FSK wave having two possible frequencies $\omega_c \pm \omega_d$ in each T-interval.

Of course it is possible to generate an FSK wave without regard to phase continuity as is implied in Eq. 1. This may be accomplished by having many oscillators tuned to the various desired frequencies and then switching their outputs in accordance with the information source. This method gives rise to undesirable switching transients in addition to complicating the transmitter and therefore is rarely used in practice particularly when the number of frequencies is large.

Spectral properties of digital FM signals have been extensively studied in the literature[2]. The general problem of understanding the relationship between baseband spectra and angle-modulated spectra is an old one and is hampered by the nonlinear relationships involved. However when the modulation is digital exact formulas for the spectral density of an FM wave are available. As would be expected because of the nonlinear nature of the modulation, FM techniques generally alter the baseband spectral shaping and can spread the bandwidth for some choice of modulation parameters. This provides the designer with the flexibility of generating large time-bandwidth product signals on one hand and on the other hand, bandwidths of the same order as double-sideband (DSB) AM can also be achieved. Since constant-envelope FM waves are generally not strictly band limited, we must adopt a practical definition of bandwidth. For instance we may regard the frequency band outside of which the intensity of the spectrum is below some arbitrarily small fraction of its peak as a practical measure of bandwidth.

For a baseband signal having L possible levels, the data symbols can be expressed as

$$a_n = 2k - (L+1); \quad k = 1,2,\ldots L \tag{4}$$

for all n (L is assumed to be even).

It can be shown that the normalized spectral density of the wave 1 is given by

$$\frac{G(\beta)}{A^2 T} = \begin{cases} \frac{1_1}{2} + \mathrm{Re}\left\{\frac{1_2^2}{1 - C_a(\omega_d T)\exp(-2\pi i\beta)}\right\}, & |C_a(\omega_d T)| < 0 \\ \frac{1_1}{2} + \left|<\frac{\sin\gamma}{\gamma}>\right|^2 \frac{1}{2\pi}\left[\sum_n \delta(\lambda - 2\pi n) - 1\right], & |C_a(\omega_d T)| = 1 \end{cases} \tag{5}$$

[2] See references (3), and (4) for bibliography.

where[+]

$$I_1 = \left\langle \left(\frac{\sin \gamma}{\gamma}\right)^2 \right\rangle$$

$$I_2 = \left\langle \frac{\sin \gamma}{\gamma} e^{-i\gamma} \right\rangle$$

$$\gamma = \frac{T}{2} (\omega - \omega_c - a_n \omega_d)$$

$$\beta = \frac{T}{2\pi} (\omega - \omega_c)$$

$$\lambda = T (\omega_c - \omega) + \pi$$

and

$$C_a(\omega_d T) = \frac{2}{L} \sum_{k=1}^{k=L/2} \cos \omega_d T (2k-1)$$

is the characteristic function of the random variable a evaluated at $\omega_d T$.

By analyzing formula (5) it can be seen that the shape of the spectrum critically depends on the single parameter ω_d T. When the magnitude of the characteristic function $C_a(x)$ evaluated at $x = \omega_d T$ is unity the spectrum contains discrete lines (delta functions) occurring at $\beta = n - \frac{1}{2}$ for all n with minimum spacing equal to the baud. Reference (4) shows various, plots of the spectral density versus normalized frequency $\beta = (\omega - \omega_c) \frac{T}{2\pi}$ for different values of $\omega_d T$ and L - the number of levels. For instance in the binary case when $\omega_d T = \pi/2$ the spectrum exhibits a very smooth roll-off with a center peak occurring at the carrier frequency. For this case the intensity is sufficiently low at 3/4 the data rate on either side of the carrier. Thus it is possible to conclude that for this choice of the parameter $\omega_d T$, a bandwidth of approximately 1.5 times the data rate is necessary to achieve reasonable performances. This in fact is born out experimentally[(5)]. From the curves given in Reference [(4)] it can also be seen that as $\omega_d T$ is increased to approximately 0.65π, the spectrum becomes approximately rectangular. Also as $\omega_d T$ approaches π the curve is smooth again but of greater spread and contains delta functions in addition to the continuous pectrum. As we shall see the performance of digital FM systems is critically dependent on the choice of the parameter $\omega_d T$.

Discrimination Detection of Digital FM A block diagram of a digital FM system is shown in Figure 1. The data source emits a sequence of multilevel symbols $\{a_n\}$ which we shall assume are independent of each other and assume the different integers with equal probability. As was discussed in the previous section, the VCO generates a signal specified by Eq. 1. In practical situations, a transmitting filter is provided whose function

+The symbol <·> denotes the ensemble average

it is to restrict the energy in the frequency-modulated wave to a range of frequencies passed by the medium, (we shall assume throughout this discussion that the transmission medium is a time invariant linear filter). It is therefore convenient to combine the medium characteristics with those of the transmitting filter into a single composite network function determining the signal presented to the receiving bandpass filter in the absence of noise.

The receiving bandpass filter is included and its function is to exclude out-of-band noise. It can also serve to shape the signal waveform and may include compensation for linear in-band distortion suffered in transmission. Usually two contradictory attributes are sought in the filter - a narrow band to reject noise and a wide band to supply a good signal wave to the detector. An opportunity for an optimum design thus exists.[6]

The frequency detector is assumed to differentiate the phase with respect to time. The post-detection filter can perform further noise rejection and shaping in the baseband range. We shall assume that the post-detection filter is matched to the baseband pulse g(t). In order to obtain estimates of the symbol sequence $\{a_n\}$, the output of the filter is periodically sampled. We thus must postulate that the receiver has independent timing information.

The noise-free input to the detector will be written in the form

$$V_r(t) = P(t) \cos(\omega_c t + \theta) - Q(t) \sin(\omega_c t + \theta) \tag{6}$$

where P(t) and Q(t) represent in-phase and quadrature signal modulation components respectively. Such a resolution can always be made, even though the details in actual examples may be burdensome. The added noise at the detector input is assumed to be gaussian distributed with zero mean and can likewise be represented as

$$V(t) = x(t) \cos(\omega_c t + \theta) - y(t) \sin(\omega_c t + \theta) \tag{7}$$

where x(t) and y(t) are independent baseband gaussian noise waves.

Let the input signal-plus noise to the limiter be

$$\begin{aligned} V_r(t) + V(t) &= x'(t) \cos \omega_c t - y'(t) \sin \omega_c t \\ &= R(t) \cos[\omega_c t + \Phi(t)] \end{aligned} \tag{8}$$

where

$$R(t) = [x'(t)]^2 + y'(t)]^2$$

and

$$\Phi(t) = \tan^{-1} y'(t)/x'(t)$$

(We have set the constant angle θ to zero without loss of generality). With these representations the output of the discriminator is taken to be

$$\frac{d\Phi}{dt} = \frac{x'(t)\,\dot{y}'(t) - y'(t)\,\dot{x}'(t)}{R^2(t)} \tag{9}$$

where the dots over x' and y' denote differentiation with respect to time. The post-detection matched filter acts on the quantity 9 and the sampled output of the matched filter is

$$q' = \int_0^T \Phi(t)\, dt \tag{10}$$

The observable q' contains the desired information and its statistical behaviour is analyzed in the next section.

The Click Theory Recently S. O. Rice [7] and previously J. Cohn [8] attacked the old threshold problem fn FM receivers by using the notion of “clicks”. It has been observed that when the noise at the input of an FM receiver is increased beyond some value, the receiver “breaks”, that is, for a given (S/N) at the input, a much poorer (S/N) at the output is measured than would be predicted from a linearized analysis of the receiver. Before the breaking point, clicks are heard in the output of an audio receiver. As the input noise is further increased, the clicks merge into a sputtering sound. Rice’s approach is to relate this breaking point to the expected number of clicks per second at the output due to the added noise at the input.

While in analog applications the criterion of (S/N) transfer is satisfactory, in digital data transmission it does not by itself furnish an adequate performance criterion. Usually performance is judged in terms of error rates which cannot be predicted from the (S/N) for nonlinear receivers. The error rate clearly depends on the statistical distribution of the output noise. In good systems, the errors are very infrequent and are associated with rare peak noise conditions. The statistical structure of the occurrence of infrequent noise peaks and the manner in which they cause errors in FM receivers will be our main concern here.

In the past several years a theory, based on the notion of “clicks” has been advanced[9],[10],[11] from which the performance of digital FM detectors with arbitrary processing gain can be predicted. In order to facilitate analysis and to gain physical

insight we assume that the medium has perfect transmission at all frequencies and the transmitting and receiving filters are inverses to one another. Because of these assumptions it is possible by use of rotating coordinates to write Eq. 10 as

$$q = q' - a_n \omega_d T = \int_0^T \frac{x(t)\,\dot{y}(t) - y(t)\,\dot{x}(t)}{(x^2(t) + y^2(t))} \tag{11}$$

where y(t) is now a zero mean quadrature gaussian noise process, while x(t) is an in-phase noise process with mean equal to the FM carrier amplitude A.

To exhibit the notion of "clicks" in digital FM reception we proceed formally with Eq. 11 and define a quantity

$$r(t) = y(t)/x(t) \tag{12}$$

and express q as a line integral

$$q = \oint_{r(0)}^{r(T)} \frac{dr(t)}{1 + r^2(t)} = \int d\phi(t) \tag{13}$$

In (13) we have written $d\phi = d(\tan^{-1} y/x)$, but of course we do not mean that ϕ is evaluated using some fixed branch of $\tan^{-1} y/x$ since this would give ϕ as a single valued function of y and x and would not allow for the fact that as we circle once about the origin in the xy-plane ϕ is increased by 2π. The noise processes y(t) and x(t) wander about the xy-plane (see Figure 2), usually staying close to their mean values but occasionally taking large excursions and encircling the origin. Each infinitesimal portion of the path contributes an amount $d\phi$ to the output and all these small amounts from all the small portions of the path must be added together to form the total contribution q. It is easy to see that q depends on the path taken, not just on its endpoints. A simple mathematical reason for this is that the transformation (12) is undefined possibility exists of stopping the process immediately after we have crossed over to the left-half plane. We will show later that for large signal-to-noise ratios, these end-effects may be neglected because they occur with a probability that is asymptotically negligible compared with the probability of a click.

An important fact to observe before proceeding with the analysis is that q can be decomposed into the sum of three random variables. The first two random variables appearing in (15) are continuous and bounded. Their probability densities are related to the elementary statistics of x(t) and y(t). The third random variable is a discrete one, whose probabilities are determined from the probabilities of zero-crossings of x(t) and y(t).

The remarks made thus far about the effect of noise on FM reception have been general; no assumptions have been made about the statistical nature of the additive disturbance. In order to obtain quantitative results some definite assumptions are necessary. We thus set ourselves the task of studying the structure of the probability distribution of q when the input noise statistics are those of a gaussian process having a symmetric Spectral density about the carrier. From these distributions we determine the error rates as a function of the pertinent system parameters. We shall make no attempt to derive an exact probability density for the random variable q. This is not a mathematically tractable problem since it requires knowledge of the probability distribution of zero-crossings of random processes. This by itself has been an area of investigation for many years without too much success. The probability distribution of the zero-crossings of most elementary random processes is not currently known.

In order to permit an analysis of the model two assumptions are made, both of which are quite reasonable. These two assumptions taken together state that the three random variables that determine q via (15) are all independent. This statement is separated into two assumptions because their individual justification stems from two different physical arguments, one having to do with bandwidth and the other with signal-to-noise ratio. The first assumption states that $\tan^{-1} r(T)$ and $\tan^{-1} r(0)$ are independent. For flat gaussian noise input this will be a good approximation if $T \geq 1/W$, where W is the input noise bandwidth. Since T is also the signaling interval, and the correlation function of the input noise has its first zero at $\tau \sim 1/W$, the motivation for this assumption is clear. The second assumption, somewhat harder to justify, states that n(T) is independent of the previous two random variables and the clicks, which comprise n(T), are independent, from one another. This is clearly an assumption expressing an intuitive feeling that the clicks occur rarely and of sufficiently short duration. In general, they will be rare if the Signal-to-noise ratio is large, and short if the bandwidth satisfies $W \geq 1/T$ as required above.

These two assumptions plus the identification of crossings of the negative x-axis by the moving point in the xy-plane (as calculated by Rice) with the occurrence of a click shall constitute the model for the error mechanism.

The Basic Probability Distributions In order to enable the calculation of error rates we collect and summarize several basic formulas which may be deduced from the click theory.

We begin by observing that if y is a zero mean gaussian variable with whenever x(t) = 0. Further, the paths taken in the xy-plane are random, and q is therefore, a random variable with some probability density related to the statistics of r(t). Unfortunately, this probability density is not determined solely by the elementary statistics of r(t). As will be

seen, in addition to the elementary statistics of r(t) the distribution of its singularities on the time axis enters the picture. The singularities of r(t) are determined by the zero-crossings of x(t). Thus, the behaviour of FM receivers is intimately related to the structure of the zero crossings of the added noise.

To see how to handle the situation, visualize the following hypothetical state of affairs. Suppose for $0 \leq t \leq T$ we have that $y(t) > 0$, and the x(t) is positive for awhile, decreases once through zero at $t = t_o$, and then remains negative. A possible plot of r(t) versus t over this time interval is then shown in Figure 3. For this particular path one has

$$q = \int_{r(0)}^{\infty} \frac{dr}{1 + r^2} + \int_{-\infty}^{r(T)} \frac{dr}{1 + r^2} = \int_{-\infty}^{\infty} \frac{dr}{1 + r^2}$$

$$+ \int_{r(0)}^{r(T)} \frac{dr}{1 + r^2} \tag{14}$$

In (14) the straightforward interpretation of the integrals is meant. Evaluating the infinite one obtains for this path

$$q = \pi + \tan^{-1} r(T) - \tan^{-1} r(0)$$

where $\tan^{-1} x$ means the principal value inverse tangent function, $|\tan^{-1} x| \leq \pi/2$. In general, one has the result that

$$q = \tan^{-1} r(T) - \tan^{-1} r(0) + n(t)\pi, \tag{15}$$

where $\tan^{-1} x$ again has the principal value interpretation and n(T) is an integer (which may be positive, negative, or zero) and is related to the number of times x(t) vanishes in the interval T and to the sign of y(t) when x(t) vanishes. For large signal-to-noise ratios it is clear that if x(t) vanishes by going to zero from the positive side that it will almost immediately be followed by another vanishing of x(t) in the other direction. If y(t) has not changed, the contribution of the "return trip" to n(T) will cancel the contribution from the previous crossing of the y-axis. On the other hand, if y(t) does change sign so as to cause an encircling of the origin then the contribution to n(t) will be the same as the previous crossing. The net contribution to n(T) of a number of paths is shown in Figure 4. The paths which have $\Delta n = \pm 2$ are immediately recognized as the "clicks" discussed by Rice. The "clicks" are not the only contribution to n(T) however. There is also a contribution because of the fact that at $t = 0$, when our process begins, we may be in the middle of a large noise fluctuation and be over in the left-half plane. Immediately afterwards, at $t = 0 +$, we will experience a contribution of ± 1 to n(T); a similar situation

may prevail at time t = T, when a

variance σ^{2+} and x is another gaussian variable with the same variance but of mean value equal to A, then the density $\rho(\bar{\Phi})$ where $\tan \bar{\Phi} = y/x$ and $\bar{\Phi}$ has the full range of 2π is well known and is given by Bennett,[12]

$$\rho(\bar{\phi}) = \frac{\exp(-\rho)}{2} + \frac{1}{2}\sqrt{\frac{\rho}{\pi}}\cos\bar{\phi}\exp(-\rho\sin^2\bar{\phi}) \qquad (16a)$$
$$\cdot[1 + \operatorname{erf}(\sqrt{\rho}\cos\bar{\phi})],$$

where $\rho = A^2/2\sigma^2$.

The probability P_L of finding the signal point in the left half of the xy-plane is

$$P_L = Pr(x<0) = \frac{1}{2}\operatorname{erfc}\frac{A}{\sqrt{2\sigma}} \sim \frac{\exp(-\rho)}{2\sqrt{\pi}\sqrt{\rho}} \qquad (16b)$$

Eq. 16 will be of use in the arguments used to discard the "end effects" at t = 0 and t = T spoken of earlier. Eq. 16 also immediately yields the probability density $\rho(\phi)$ for $\phi = \tan^{-1}(y/x)$, $\rho(\phi) = \bar{\rho}(\phi) + \bar{\rho}(\phi + \pi)$

$$= \frac{\exp(-\rho)}{\pi} + \frac{\sqrt{\rho}}{\pi}\cos\phi\exp(-\rho\sin^2\phi)\operatorname{erf}(\sqrt{\rho}\cos\phi) \qquad (17)$$

for $|\phi| \le \pi/2$.

Suppose ϕ_1 and ϕ_2 are two independent anqles which have the density (17), and define an angle $\Phi = \phi_1 - \phi_2$ $|\Phi| \leq \pi$. It will be of interest for us to determine the probability P_ϕ that Φ exceeds some angle $\phi > 0$, i.e., we would like to determine

$$P_\phi = \int_{(-\pi/2)}^{(\pi/2)-\phi} d\phi_2 \int_{\phi_2+\phi}^{\pi/2} d\phi_1\, \rho(\phi_1)\,\rho(\phi_2)\,\rho(\phi_2), \quad \phi > 0. \qquad (18)$$

The asymptotic evaluation of (18) is carried out in detail in reference (11); we distinguish three cases:‡

Case I; $0 < \phi < \pi/2$

$$P_\phi \quad \frac{1}{\sqrt{8\pi}}\frac{\cot(\phi/2)\exp[-2\sin^2(\phi/2)]}{\sqrt{\cos\phi}\sqrt{\rho}} \qquad (19a)$$

+ Recall that even though our x(t) and y(t) are not independent processes because the noise spectrum will not be symmetrical about $(\omega_o + a_n \omega_d)$, they are independent variables.

‡ In (19) the symbol "~" is used to denote asymptotic equality; this has also been used in (16). Also (19a) and (19c) do not hold if ϕ gets too close to the end points of the appropriate interval. As a rough rule ϕ should not be closet than $1/\sqrt{\rho}$ radians to the end points.

Case II; < = $\pi/2$:

$$P_\phi \sim \frac{1}{4} \exp\ (-\rho). \tag{19b}$$

Case III; $\phi > \pi/2$:

$$P_\phi \sim \frac{\exp\ [-\rho(1 + \cos^2 \phi)]}{2\pi\sqrt{\pi\rho}\sqrt{\rho \sin \phi \cos^2 \phi}}. \tag{19c}$$

The most important characteristic of the result (19) is the dependence of the exponent on angle, since for large p the nonexponential factors are relatively slowly varying.

The final item that we discuss in this section is the density of n(T), or rather we discuss the density of that part of n(T) that arises from the clicks ($\Delta n = \pm 2$), ignoring $\Delta n = \pm 1$ contributions. For this we need only take over some ideas and formulas from Rice. We have that (ignoring $\Delta n = \pm 1$)

$$\pi n(T) = 2\pi N(T), \tag{20}$$

where N(T) is the number of clicks that occur in time T. Following Rice, we assume that all clicks are independent and that those tending to increase (decrease) ϕ by 2π form a Poisson process with rate of occurrence N + (N-). In general, in the presence of modulated signal, N_+ and N_- are not equal. The probability density $\rho(z)$ of z = N(T) is then given by

$$\rho(z) = \exp\ [-(N_+ + N_-)T] \sum_{k=-\infty}^{\infty} \delta\ (z - k) \left|\frac{N_+}{N_-}\right|^{k/2} \cdot I_k\ (2T\sqrt{N_+ N_-}); \tag{21}$$

as may be shown by forming the discrete convolution of the densities of the positive and negative clicks. In (21) $\delta(\cdot)$ is the Dirac delta function and $I_k(\mu)$ is the modified Bessel function of integer order k, behaving for small μ as

$$I_k(\mu) \underset{\mu \to 0}{\to} \left(\frac{\mu}{2}\right)^{|k|} \frac{1}{|k|\ !} \tag{22}$$

also

$$I_{-k}(z) = I_k(z).$$

The type of modulation that we are concerned with is when the instantaneous frequency deviates by ω_d. from the carrier for a time T, T being the signal-average rates N_+ and N_- when the noise at the receiver input is gaussian

$$N_+ = \frac{1}{2}\left\{\sqrt{r^2 + f_d^2}\left[1 - \operatorname{erf}\sqrt{\rho + \rho f_d^2/r^2}\right]\right.$$

$$\left. - f_d \exp(-\rho)\left[1 - \operatorname{erf}\left(f_d\sqrt{\rho}/r\right)\right]\right\}$$

where $f_d = \omega_d/2\pi$

and

$$N_- = N_+ + f_d \exp(-\rho), \tag{23}$$

where [+]

$$r = (1/2\pi)(\dot{\sigma}/\sigma)$$

$$\sigma^2 = \operatorname{var} x = \operatorname{var} y$$

$$\dot{\sigma}^2 = \operatorname{var} \dot{x} = \operatorname{var} \dot{y}.$$

$$N_+ \sim \frac{1}{4\sqrt{\pi}} \frac{1}{\rho^{3/2}} \frac{r}{\left[1+(f_d^2/r^2)\right](f_d^2/r^2)} \exp\left[-\rho(1 + f_d^2/r^2)\right]$$

$$N_- \sim N_+ + f_d \exp(-\rho). \tag{24}$$

Thus, we see that for large ρ an ever greater majority of clicks occur in the negative direction ($f_d > 0$) and for our purposes of computing error rate, the clicks in the positive direction may be neglected; i.e., we shall use

$$N_+ \sim 0$$

$$\text{for } f_d > 0. \tag{25}$$

$$N_- \sim f_d \exp(-\rho)$$

For $f_d < 0$ the situation is reversed of course. We note that the effect of the clicks on a modulated carrier is to tend to make the measured frequencies appear closer to the carrier

---

+ The case ($\omega_d = 0$ corresponds to no modulation. Also, for ease of writing, we no longer explicitly consider the factor a_k.

frequency than the transmitted frequencies. That is, confining oneself for the moment to only errors caused by clicks, frequencies transmitted higher (lower) than the carrier will be measured to be at that frequency or a lower (higher) one, when the noise is small.

Since we shall use approximation (25), the distribution (21) for z = N(T) may be replaced by the simpler Poisson one, where the probability of getting exactly N (negative) clicks in time T is given by+

$$N(T) = \frac{\exp(-N\_T)(N\_T)^{N(T)}}{[N(T)]!} . \tag{26}$$

Also the probability of getting M or more clicks is, for large signal-to-noise ratios, approximately the probability of getting exactly M clicks.

Distribution of Output and Probability of Error Eq. 16, 19, 25, and 26 provide the information required to calculate the distribution of q, (Eq. 15). In principle we simply convolve the continuous density of $[\tan^{-1} r(T) - \tan^{-1} r(0)]$ with the discrete density of $n(T)\pi$. In Figure 5, we give a qualitative sketch of the result, neglecting end effects. This picture is intended to show that the density consists of a central lobe about the transmitted frequency extending to $\pm \pi$ on each side, which is the density of $[\tan^{-1} r(T) - \tan^{-1} r(0)]$, plus identically shaped lobes displaced by intergral multiples of 2π toward lower frequencies ($f_d > 0$). These displaced lobes are weighted by the probability of getting tge appropriate number of clicks to effect the displacement. Thus, the lobe occupying the space $-2n\pi \pm \pi$ is weighted by the probability of getting exactly n clicks in time T. For n = 0 the weighting is essentially one, for large S/N. There are, strictly speaking, similar lobes and weightings on the opposite side as well, but these weights are, for large S/N, negligible compared to the corresponding lobe we have drawn. That is to say, the first lobe on the right (not shown in Figure 5) has small probability compared to the first lobe on the left, but has a large probability compared to the second lobe on the left. Nevertheless, we have neglected to include it because we will generally be concerned with probabilities like $\Pr[|q - f_d T| > \phi]$, and thus corresponding weights are important. We dwell on this point because it is conceivable that for some practical or conceptual application the neglect would not be justified.

The discussion given above is still not quite complete it is modified when we include end effects. The principal correction that inclusion of end effects will cause is to add two more side lobes, one over the interval L-27r,01 and the other over the interval [0,2Tr]. The weightings of these lobes certainly should not exceed the estimate given in (16), and this will be enough to exclude them for our purposes.

+ We confine ourselves to $f_d > 0$. Exactly analogous consideration apply to $f_d < 0$. The case $f_d = 0$ occurs if an odd number of frequencies are allowed.

We now apply our results to some typical calculations. Consider the case of narrow band[+] FM (defined by $\Delta f_d T < \pi$), where one has J equally spaced frequencies of separation $f\Delta_d$ crowded into a bandwidth W. The probability of error for any one of the frequencies[++] (not situated at the ends) is the area outside of the interval bounded by lines L_2 and L_3 in Figure 5. If L_2 and L_3 are defined by $|q|$ then the probability of error for such a frequency is from (19a).

$$P_e = \frac{1}{2\pi} \frac{\cot \Phi/2 \exp [-2\rho \sin^2 (\Phi/2)]}{\sqrt{\cos \Phi}\sqrt{\rho}} \tag{27}$$

where, if one assumes that the bandwidth $W = J \Delta f_d$ one would take

$$\Phi = \frac{\pi WT}{J} . \tag{28}$$

Our requirement that $\Delta f_d T < \pi$ implies $J > 2$ for the narrow-band formula to be applicable (assuming $WT = 1$). Note $\sin^2 (\Phi/2)$ is less than $\frac{1}{2}$, and thus the exponent in (27) is $\exp [-k\rho]$, where $k < 1$. Now the contribution of the clicks to P_e is essentially the area A_L of the first side lobe which is by (25) and (26)

$$A_L = f_d T \exp (-\rho). \tag{29}$$

But expression (29) is, asymptotically, exponentially small compared to (27). Likewise, the area due to the side lobes caused by end effects is exponentially small, and the probability of error for narrow-band FM is given by (27).

Next, consider the asymptotic evaluation of$_e$ P for the case of orthogonal signals; this case corresponds to $(\Delta\omega_d)T = \pi$, we assume that the thresholds are spaced midway between the frequencies. Thus, (for a frequency not on the edges) we have that the error probability is given by the area outside of that bounded between the lines L_1 and L_4. The contribution from the major lobe is, from (19b),

$$\frac{1}{2} \exp (-\rho).$$

In addition, the area of the first side lobe is asymptotically comparable to this and is

$$f_d T \exp (-\rho),$$

being weakly dependent on the frequency sent. In fact, for the nth signal (j = 2n) we have for orthogonal signals that

---

+ Note the special sense in which the term is used here.

++ The P_e for a frequency at the end is one-half the expression (27).

$$f_d T = \frac{n}{2}, \quad n = 1, 2, \ldots, \frac{J}{2} .$$

The average error rate is then, for orthogonal signals (J of them, J even, and equally spaced signals and thresholds),

$$P_e = \left(\frac{1}{2}\right) \exp\left[-\rho\right] + \left(\frac{1}{4}\right) (J/2 + 1) \exp(-\rho). \tag{30}$$

Eq. 30 is indeed a surprising result. The first term of (30) is the probability of confusing the transmitted frequency with one of its nearest neighbors. The second term is the (average) probability of confusing it with its second nearest neighbor closest to the carrier. This is because the area from $(-\pi)$ to $(-3\,\pi/2)$ is, by application of (19b), negligible compared to the area from $(-3\,\pi/2)$ to $(-5\,\pi/2)$. Thus, it states that for the multilevel situations considered here (a not unreasonable one) it is to confuse it with a particular one of its second nearest neighbors. We see from (30) that the error rate from the continuous part of the output is comparable to the error rate caused by clicks.

For a final example, consider the wide-band situation where the signals are loosely packed in the band; i.e., $(\Delta\omega_d)\,T > \pi$. Now no errors will be caused by the continuous part of the output; only clicks will cause errors. If the frequencies are widely spaced a single click may not cause an error; several clicks during the time interval T may be required. Thus, suppose that the frequencies are spaced so that the phase differences of nearest neighbors is $(\Delta\omega_d)\,T = 2n\,\pi$, n being any positive integer. The probability of error will then be the probability of getting n (or more) clicks in time T, which from (25) and (26) behaves as

$$\frac{(f_d T)^2 \exp(-n\rho)}{n!} \geqq \frac{(n/2)^n \exp(-n\rho)}{n!} \geqq \frac{1}{2} \exp(-n\rho) \tag{31}$$

The coefficient in (31) is at least as bad as for the orthogonal case, but the important item is the exponent. Superficially at least it appears that we have gained in performance by spacing the frequencies widely, since the exponential has changed from $e^{-\rho}$ from the minimum orthogonal case $(\Delta\omega_d\text{-}T = \pi$. to $e^{-n\rho}$. It must be realized, however, that here we are talking about different ρ's. The bandwidth for the case under consideration is essentially 2n times the minimum orthogonal one and therefore, for the same signal power, the nominal value of ρ has decreased by 2n, and one has in fact not gained a factor of n in the exponent. In addition to the bandwidth penalty, error performance has actually suffered too.

Comparison with Optimum It is possible to demonstrate how the FM discriminator compares with the ideal detector when used to detect orthogonal signals; i.e. when $\Delta\omega_d T = \pi$. It is known that when optimum detection is used for any orthogonal set of signals, the (exponential part of the) error rate behaves as exp $[-E/N_o]$, where E is the signal energy (assumed common to all J levels) and $N_o/2$ is the (two-sided) spectral density of the noise. If we let S denote the average signal power, such that E = ST, and estimate the total bandwidth W for large J by $W \sim J/(2T)$, we see that the ideal exponent becomes exp $[-J\rho/2]$. However, we had seen that, regardless of the number of levels, the discriminator error rate $\Delta\omega_d T = \pi$. behaves as exp $[-\rho]$. Thus, we have lost a factor of J in the error exponent by substituting discrimination detection for matched filter detection.

An important conclusion may immediately be drawn concerning the performance of conventional FM receivers as detectors of orthogonal signals. Our results show that the receiver is indeed inferior in performance when compared with the optimum. This fact has been stated by Wozencraft and Jacobs [13] and the reasons are clear from this analysis. The FM receiver admits too much noise at its front-end which cannot be cleaned by the post-detection filter because of the nonlinear anomalies, namely the clicks. As a matter of fact, the amount of noise grows in direct proportion to the number of orthogonal signals, hence the inferior exponent. On the other hand, the optimum detector is composed of a bank of matched filters. The noise power at the output of each filter does not grow with the number of signals; it is a fixed constant determined by the bandwidth of the filter, which roughly needs be no broader than the symbol rate, 1/T.

This poor performance of conventional FM receivers when used to detect data might be remedie by employing an FM with feedback such as described in references [14] and [15]. The physical argument to support this contention is often stated as follows. In the absence of the feedback loop, the IF filter must be wide enough to pass the total swing of the incoming signal. However, since the feedback loop tracks the incoming frequency, this IF filter, whose width determines the noise variance, could be narrowed and less noise would be admitted.

This possibility of making use of FM with feedback to im rove the error rate in digital systems has been suggested in reference (13). Unfortunately a mathematical treatment of this difficult problem does not exist at present.

Signaling over the Gaussian Dispersive Medium When a dispersive medium is present between transmitter and receiver, error rate calculations become extremely complicated and can only be carried out in certain special cases. Although the mechanism of error production can still be explained in terms of the clicks, very few quantitative results are available.[16] The difficulty arises because the signal-plus-noise is subjected to a mixture of linear and nonlinear operations. The dispersive medium

introduces intersymbol interference in a very complicated manner affecting the rate at which clicks are produced.

A major source of difficulty in the analysis of digital FM detection is the presence of the post-detection filter. If one assumes that this filter is ideal, in the sense that it does not perform significant baseband processing, the mathematical problems become tractable. In many applications this assumption is valid. In particular it is a good one when the modulated signal bandwidth is of the same order as that of the baseband signal. In these applications exact formulas for the probability distribution of the instantaneous frequency of signal-plus-noise are available in references (17) and (18). Even here the use of these formulas becomes very tedious since it is necessary to make separate calculations for every input symbol sequence. This often can only be accomplished with the aid of a high speed digital computer. Only the binary case has been adequately treated in the literature(19), (20), (21). Experimental results with multilevel FM operating over a moderately dispersive medium have been reported on in the literature.(5) No adequate theory however is at present available capable of handling situations where the post-detection filter cannot be neglected.

References

1) Powers, K. H., Bedrosian, E., “The Analytic Signal representation of Modulated Waveforms” Proc. IRE, Vol. 50, October 1962.

2) Mazo, J. E., SaIz, J., “Spectral properties of Signal-Side-band Angle Modulation” IEEE Trans. Comm. Tech., February 1968.

3) Bennett, R. W., Davey, J. R., “Data Transmission” McGraw-Hill Book Co., 1965.

4) Lucky, R. W., SaIz, J., Weldon, E. J., Jr., “Principles of Data Communication” McGraw-Hill Book Co., 1968.

5) Salz, J., Koll, V. G., “An Experimental Digital Multilevel FM Modem,” IEEE Trans., COM-14, 1966, pp. 259-265.

6) Salz, J., “Performance of Multilevel Narrow-Band FM Digital Communication Systems,” IEEE Trans., COM-13, 1965, pp. 420-424.

7) Rice, S. O., “Noise in FM Receivers,” Chapter 25 of Times Series Analysis, Rosenblatt, M., (ed.), John Wiley & Sons, Inc., New York, 1963.

8) Cohn, John, Proc N.E.C., (Chicago) 12, 1956, pp. 221-236.

9) Schilling, D. C., Hoffman, E., and Nelson, E. A., “Error Rates for Digital Signals Demodulated by an FM Discriminator,” IEEE Trans. Communication Technology, COM-15 (August 1967), PP. 507-517.

10) Klapper, J., “Demodulator Threshold Performance and Error Rates in Angle Modulated Digital Signals,” RCA Review 27 (June 1966), pp. 226-244.

11) Mazo, J. E. and Salz, J., “Theory of Error Rates for Digital FM, BSTJ, 45(November 1966), pp. 1511-1535.

12) Bennett, W. R., “Methods of Solving Noise Problems,” Proc. IRE, 44, 1956, pp. 609-638, See (253).

13) Wozencraft, J. M. and Jacobs, I. M., “Principles of Communication Engineering,” John Wiley & Sons, Inc., New York, 1965.

14) Chaffe, J. G., “The Application of Negative Feedback to Frequency Modulation Systems, B.S.T.J., 18, July, 1939, pp. 403-437.

15) Enloe, L. H., “Decreasing the Threshold in FM by Frequency Feedback,” Proc. IRE, 50, 1962, pp. 18-30.

16) Schneider, H. L., “Click Comparison of Digital and Matched Filter Reception,” B.S.T.J., March 1968.

17) Salz, J. and Stein, S., “Distribution of Instantaneous Frequency for Signal Plus Noise,” IEEE Trans., IT-10, 1964, pp. 272-274.

18) Mazo, J. E. and Salz, J., “Probability of Error for Quadratic Detectors,” B.S.T.J., 44, November, 1965, pp. 2165-2186.

19) Bennett, W. R. and Salz, J., “Binary Data Transmission by FM Over a Real Channel,” B.S.T.J., 42, 1963, pp. 2387-2426.

20) Meyerhoff, A. A. and Mazer, W. M., “Optimum Binary FM Reception Using Discriminator Detection and I-F Shaping,” RCA Review, 22, Dec. 1961, pp. 698-728.

21) Smith, E. F., “Attainable Error Probabilities in Demodulation of Rand om Binary PCM/FM Waveforms,” IRE Trans. on Space Elec. and Telemetry, SET-8, Dec. 1962, pp. 290-297.

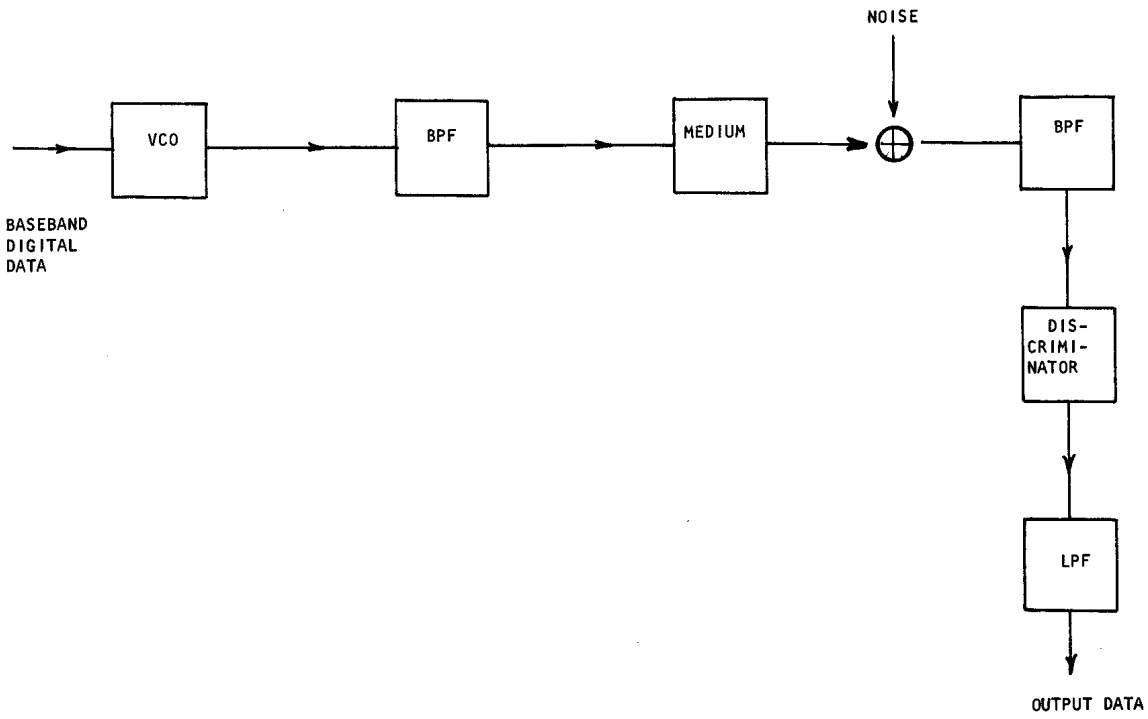

FIGURE 1 - FM DATA TRANSMISSION SYSTEM

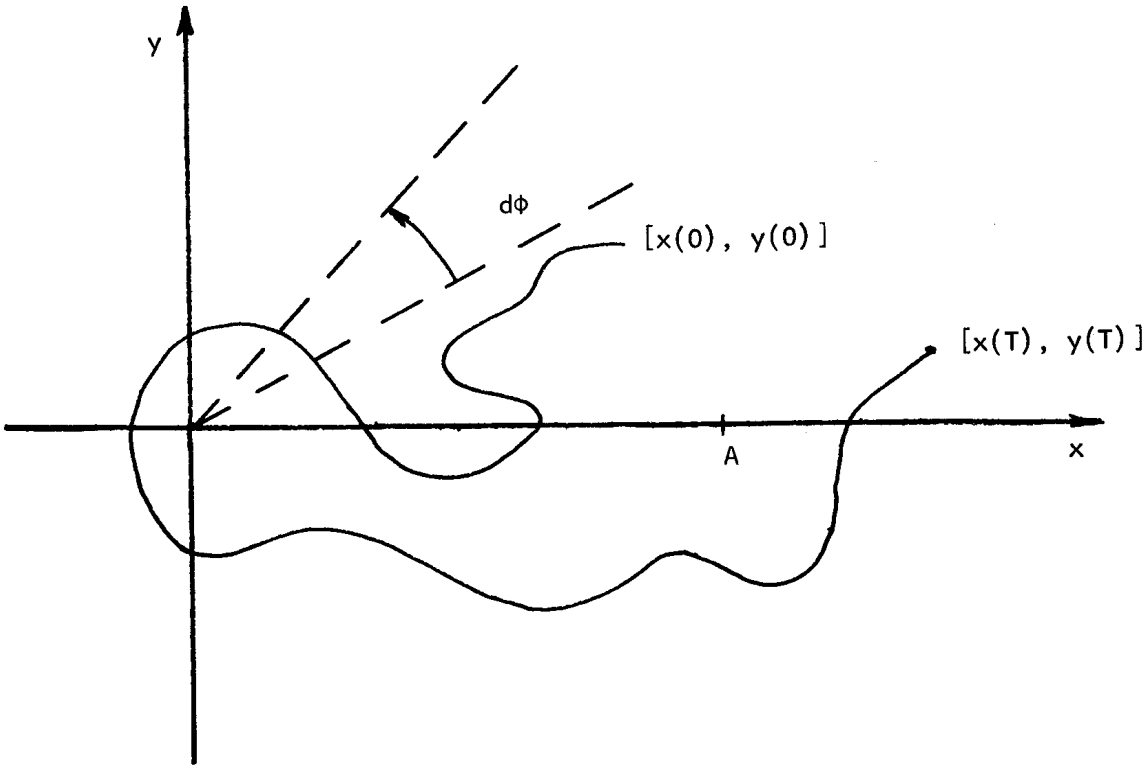

FIGURE 2 - A POSSIBLE PATH IN THE XY-PLANE TRACED BY THE NOISE FROM t=0 TO t=T

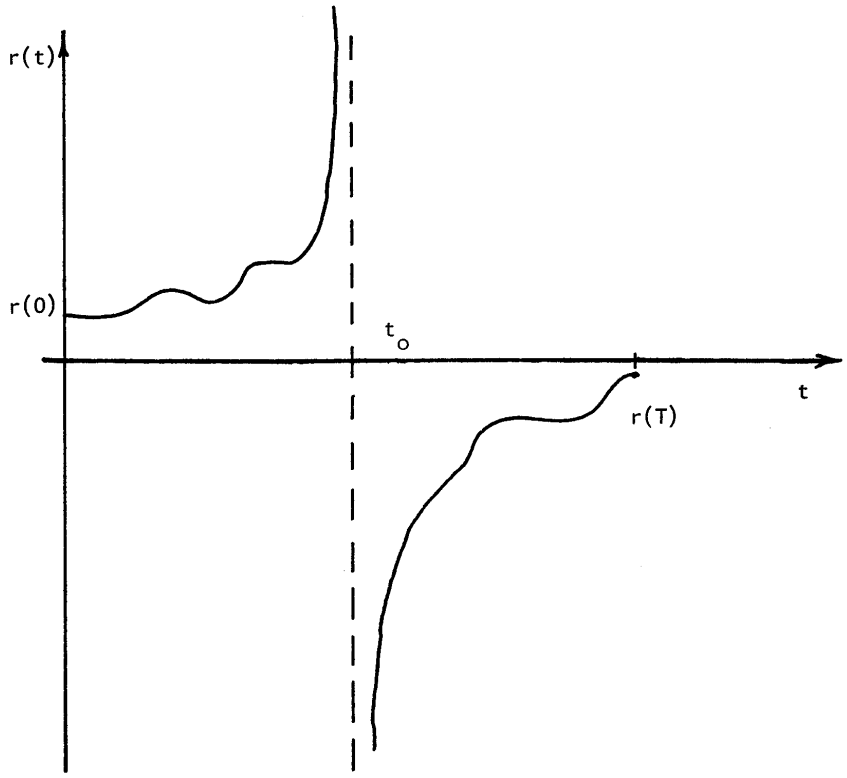

FIGURE 3 - A POSSIBLE SAMPLE FUNCTION OF r(t)

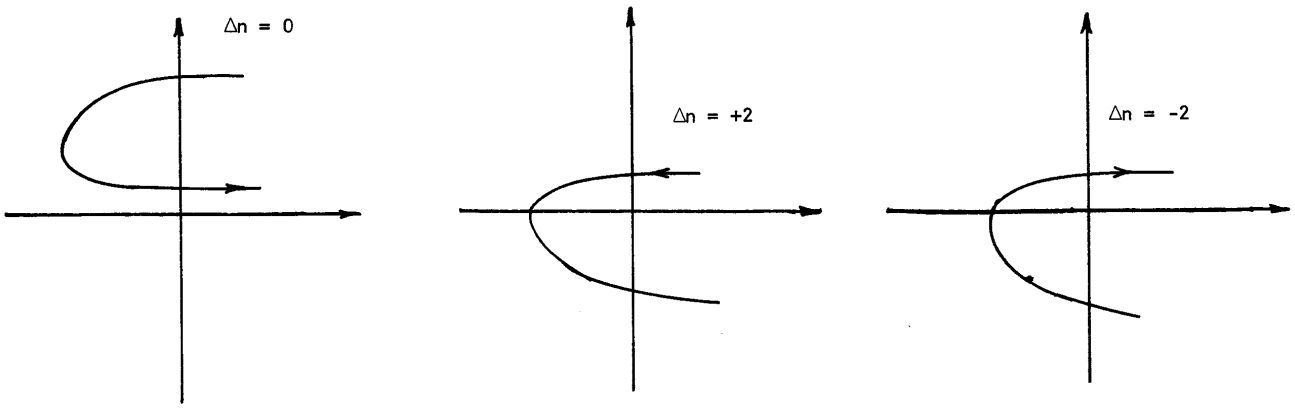

FIGURE 4 - NET CHANGES Δn IN n(T) CAUSED BY SOME TYPICAL PATHS IN THE XY-PLANE

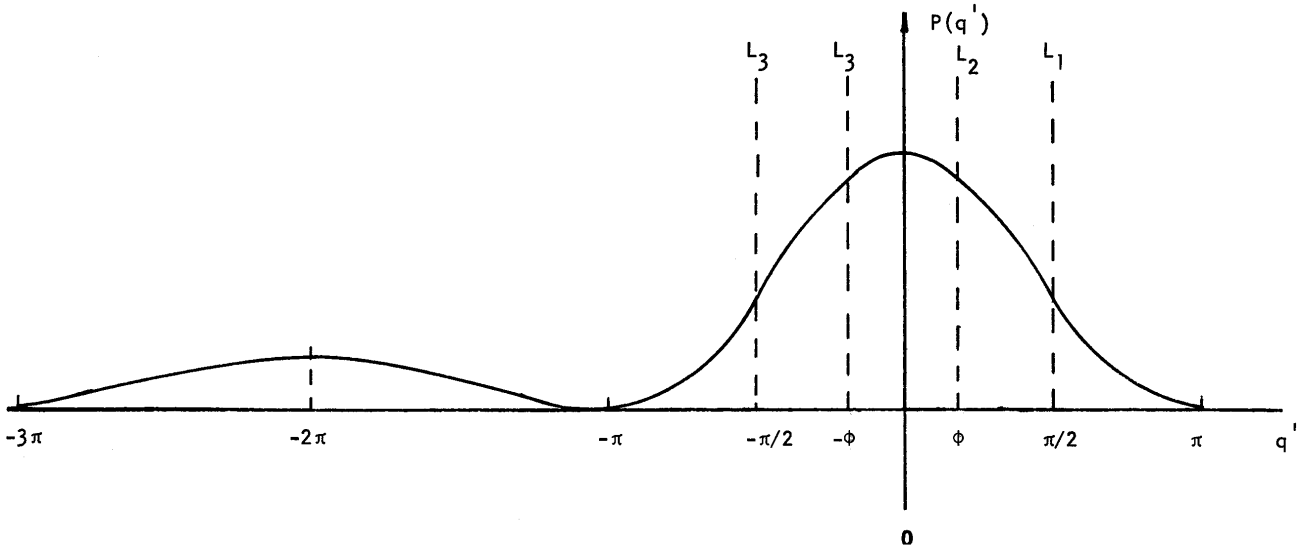

FIGURE 5 - QUALITATIVE SKETCH OF THE PROBABILITY DENSITY OF q