illustration with a transmitting antenna with effective area A_t will increase the received power by the ratio $A_t/A_{\text{isotr.}}$, and we obtain

$$P_r/P_t = A_r A_t/4\pi d^2 A_{\text{isotr.}} \tag{10}$$

Introducing the effective area (6) for the isotropic antenna, we have (1).

Limitations of Transmission Formula (1)

In deriving (1), a plane wave front was assumed at the distance d. Formula (1), therefore, should not be used when d is small. W. D. Lewis, of these Laboratories, has made a theoretical study of transmission between large antennas of equal areas with plane phase fronts at their apertures and he finds that (1) is correct to within a few per cent when

$$d \geqq 2a^2/\lambda \tag{11}$$

where a is the largest linear dimension of either of the antennas.

Formula (1) applies to free space only, a condition which designers of microwave circuits seek to approximate. Application of the formula to other conditions may require corrections for the effect of the "ground," and for absorption in the transmission medium, which are beyond the scope of this note.

The advantage of (1) over other formulations is that, fortunately, it has no numerical coefficients. It is so simple that it may be memorized easily. Almost 7 years of intensive use has proved its utility in transmission calculations involving wavelengths up to several meters, and it may become useful also at longer wavelengths. It is suggested that radio engineers hereafter give the radiation from a transmitting antenna in terms of the power flow per unit area which is equal to $P_t A_t/\lambda^2 d^2$, instead of giving the field strength in volts per meter. It is also suggested that an antenna be characterized by its effective area, instead of by its power gain or radiation resistance.[6] The ratio of the effective area to the actual area of the aperture of an antenna is also of importance in antenna design, since it gives an indication of how efficiently the antenna is utilizing the physical space it occupies.

[6] The directional pattern, which has not been discussed in this note, is, of course, always an important characteristic of an antenna

Nonlinearity in Frequency-Modulation Radio Systems Due to Multipath Propagation*

S. T. MEYERS†

Summary—**A theoretical study is made to determine the effects of multipath propagation on over-all transmission characteristics in frequency-modulation radio circuits. The analysis covers a simplified case where the transmitted carrier is frequency-modulated by a single modulating frequency and is propagated over two paths having relative delay and amplitude differences. Equations are derived for the receiver output in terms of the transmitter input for fundamental and harmonics of the modulating frequency. Curves are plotted and discussed for various values of relative carrier- and signal-frequency phase shift and relative amplitude difference of the received waves.**

The results show that a special kind of amplitude nonlinearity is produced in the input-output characteristics of an over-all frequency-modulation radio system. Under certain conditions, sudden changes in output-signal amplitude accompany the passage of the input-signal amplitude through certain critical values. Transmission irregularities of this type are proposed as a possible explanation of so-called "volume bursts" sometimes encountered in frequency-modulation radio circuits. In general, it appears that amplitude and frequency distortion are most severe where the relative delay between paths is large and the amplitude difference is small.

THE INFLUENCE of multipath propagation on the transmission properties of frequency-modulation radio circuits is of considerable interest. The subject has been treated at some length in previous papers from both experimental and theoretical standpoints.[1,2,3] It is the purpose here to extend the theoretical side in an effort to obtain a clearer understanding of the true nature of the over-all circuit transmission changes induced by multipath propagation. Experimental support of the conclusions has not been obtained, due to the lack of time and facilities brought on by the pressure of war work.

Many of the causes of multiple paths over which radio waves sometimes travel from transmitter to receiver are well known and need not be recounted here. It is sufficient to state that when these paths exist simultaneously and are of different lengths and time of travel, interference at the receiver takes place between arriving waves. This interference is manifest by alterations in the amplitude and phase-versus-frequency characteristics of the resultant received wave as compared to the wave which is transmitted. In all types of radio systems such alterations in the received wave usually result in a

* Decimal classification: R630.11. Original manuscript received by the Institute, November 13, 1945.

† Bell Telephone Laboratories, Inc., 180 Varick St., New York 14, N. Y.

[1] Murray G. Crosby, "Frequency-modulation-propagation characteristics," Proc. I.R.E., vol. 24, pp. 898–913; June, 1936.

[2] Murray G. Crosby, "Observations of frequency-modulation propagation on 26 megacycles," Proc. I.R.E., vol. 29, pp. 398–403; July, 1941.

[3] Murlan S. Corrington, "Frequency-modulation distortion caused by multipath transmission," Proc. I.R.E., vol. 33, pp. 878–891; December, 1945.

demodulated output at the receiver which is a distorted copy of the input to the transmitter.

By considering the radio system as a whole with the transmitter, transmission medium, and receiver in tandem it is possible to state the over-all transmission performance in terms similar to those used in describing the performance of an amplifier or any other self-contained transmission unit. Such performance is usually given in terms of input-output frequency characteristic or input-output amplitude characteristic or both. Any modulation produced within the transmission unit and appearing in the output is usually expressed in per cent of the fundamental output amplitude and plotted as a function of either the output or input fundamental amplitude. Characteristics such as these form a convenient picture and are adopted here to describe the over-all performance of a frequency-modulation radio system having multipath propagation between transmitter and receiver.

Because of the complexity of the analysis, consideration is given only to the case where the desired wave arrives at the receiver over two paths, each having constant time delay and attenuation over the transmitted radio-frequency band. The transmitted wave is assumed to consist of a sinusoidal carrier frequency-modulated by a single sinusoidal modulating frequency. The limiter in the receiver is assumed to have a zero-order characteristic[4] to insure limiting at low amplitudes. Noise is neglected, although in any practical consideration it becomes the controlling factor where multipath interference reduces the signal strength at the receiver below that of noise.

If two frequency-modulation waves reach the receiver over the two paths just described, they will have relative amplitude and phase differences proportional to the amplitude and phase differences between the paths. Mathematically, on applying these two waves to a frequency-modulation receiver, equations for the resultant demodulated output can be derived in conventional manner (see Appendix). These equations can be expanded so that the receiver output may be expressed in terms of transmitter input for the fundamental and harmonics of the modulating frequency. The final results as derived in the appendix are the basis of the discussion which follows.

What are called single-frequency amplitude characteristics of an over-all radio system under the influence of two-path propagation are shown in Fig. 1. With the receiver output plotted in terms of transmitter input, these curves are computed from the fundamental term of (10) in conjunction with the computed curves for D_1, in Fig. 10 in the appendix. The constants associated with the curves of Fig. 1 are

β = relative carrier phase shift between paths
α = relative modulating-frequency phase shift between paths; i.e., the relative difference in phase-shift differential between the carrier and the first-order sideband in each path
r = voltage ratio of the later wave to the earlier wave or the amplitude ratio corresponding to the loss of the longer to the shorter path
f_d/f_p = deviation index
f_d = maximum carrier-frequency deviation
f_p = modulating frequency
h = factor proportional to the signal amplitude of the transmitter input.

The co-ordinates in Fig. 1 are expressed in decibels. The signal input to the transmitter is proportional to h

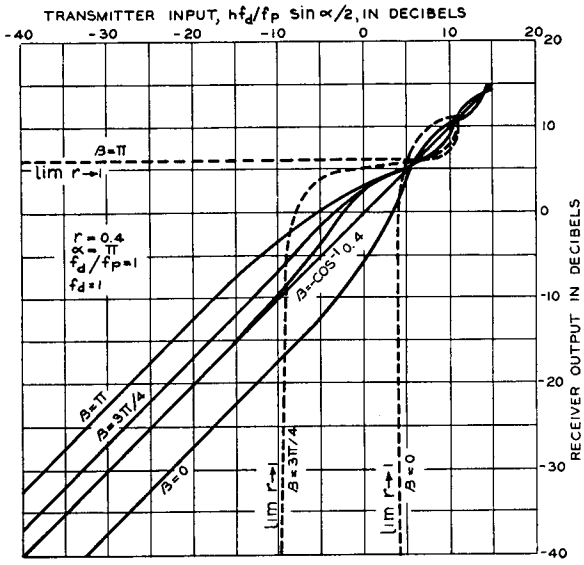

Fig. 1—Single-frequency amplitude characteristics of an over-all frequency-modulation radio system.

and is plotted as abscissa. The deviation index f_d/f_p and $\sin \alpha/2$, the sine of half the relative signal-frequency phase shift, serve as multiplying factors of h. For the curves as they are shown these two factors are unity or zero decibels. Other values are discussed later. An arbitrary reference has been chosen where $hf_d/f_p \sin \alpha/2 = 1$. Expressed in decibels this is zero on the abscissa of Fig. 1. The receiver output is given by the fundamental term of (10) and is plotted as ordinate for unit value[5] of f_d. The ordinate scale is the same as that of the abscissa so that if the over-all system is distortionless the input-output characteristic will be a straight line having a 45-degree slope. Any deviation from this slope means the output does not follow the input linearly, and amplitude distortion results.

A reference characteristic is established in Fig. 1 representing a normal distortionless system in which multipath interference does not exist. This is the 45-degree

[4] The instantaneous current-voltage relationship in a resistance limiter may be expressed $I = K|E|^n$, $n < 1$, where the sign of I follows the sign of E. A zero-order characteristic obtains when $n = 0$, and

$$I = K, \quad E > 0$$
$$= -K, \quad E < 0.$$

[5] To avoid confusion concerning the meaning of f_d it will be assumed here that it is a fixed quantity representing maximum carrier-frequency deviation in the usual sense when $h = 1$.

line passing through the point 0, 0 which is the plot of (10) when r, the amplitude ratio of the longer to the shorter path, is zero. The other characteristics shown in solid lines are plotted for $r=0.4$ in addition to $\alpha=\pi$ and $f_d/f_p=1$ mentioned previously. The parameter is the relative carrier-frequency phase shift $\beta=0, 3\pi/4, -\cos^{-1}0.4$ and π. The values of $\beta=0$ and π represent extreme deviations of the amplitude characteristic from that of a normal distortionless system. For values of β between 0 and π the deviation from the normal system becomes less as shown by the other two values[6] of $\beta=3\pi/4$ and $-\cos^{-1}0.4$.

It will be noticed that at low values of h all the characteristics approach straight lines having a 45-degree slope. This signifies that the system input and output in each case is approaching linear relationship and distortion is approaching zero as $h\doteq 0$. The displacement of the various characteristics above and below that for a normal system signifies a corresponding increase or decrease in over-all circuit transmission. This is reasonable since an inspection of the frequency spectrum of a frequency-modulation wave shows that when $\beta=0$ and $\alpha=\pi$, the separate-path carrier frequencies add while the sidebands subtract, giving a net reduction in instantaneous phase of the carrier. This, in turn, results in reduced output signal for a given signal input. Conversely, when $\beta=\pi$, and $\alpha=\pi$ the carrier frequencies subtract while the sidebands add, giving a net increase in instantaneous carrier phase resulting in an increase in output signal for a given signal input.

When $\beta=-\cos^{-1}r$ the load characteristic, at small values of h, approaches the normal as h decreases. This value of β forms, in this region, the dividing line between characteristics of increased and decreased over-all system transmission. In Fig. 1 such a characteristic is obtained when $\beta=-\cos^{-1}0.4$.

As h increases beyond these small values just mentioned, the various characteristics sooner or later depart from the 45-degree slope giving rise to amplitude distortion. The characteristic for $\beta=\pi$ up to about +5 decibels on the co-ordinates resembles somewhat the sort of overload characteristic obtained in an ordinary vacuum-tube circuit when it is driven into the region of compression. However, as h increases further, the slope of the characteristic reverses and the curve returns to cross the normal again and to continue on in undulating fashion in ever-decreasing deviations about the normal, approaching the normal as a limit.

The characteristic for $\beta=0$ up to about +5 decibels on the co-ordinates has somewhat the inverse characteristic of the one just discussed. It is similar to the load characteristic of a vacuum-tube circuit in which there is expansion associated with threshold. Such a characteristic might be obtained in a class AB amplifier. As h increases beyond the +5-decibel region, this characteristic likewise reverses in curvature and progresses in undulating fashion in decreasing deviations about the normal and approaches the normal, as a limit.

A sort of critical region seems to exist, in the neighborhood of +5 decibels, above which all characteristics for all values of r, β, and α converge in undulating fashion on the normal as h increases. This is an interesting property since the effects of fading would be greatly reduced in a frequency-modulation radio system by maintaining the useful range of h above this critical region.

It is of interest to note that the normal distortionless system characteristic (the straight line passing through 0, 0) is also the system characteristic for the two-path case when $\alpha=0$. This may be seen from (9) and (10) in the appendix when $\sin\alpha/2=0$. If the delay differential of the two paths of propagation is small enough, $\sin\alpha/2$ may be neglected and a substantially distortionless system may be obtained even with two-path propagation.

As mentioned previously, the characteristics in solid lines in Fig. 1 were computed for $r=0.4$. As the two paths of propagation approach equal transmission, the over-all system characteristics deviate further from the normal; as $r\doteq 1$ the curves for $\beta=0$, $3\pi/4$, and π approach the dotted lines as limits. The manner in which this takes place will be made clearer by comparing Fig. 1 with Fig. 2, where similar characteristics are plotted for $r=0.8$ for the same value of β.

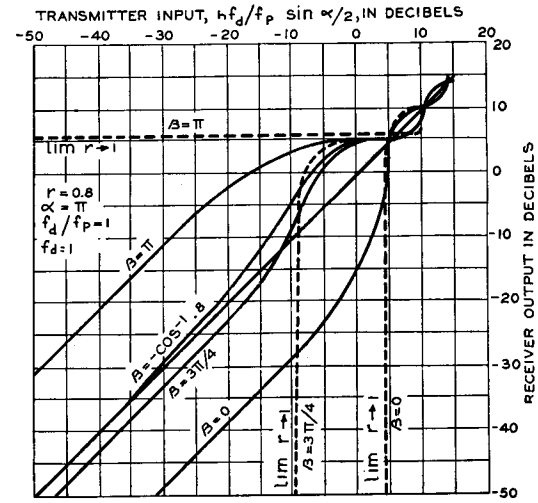

Fig. 2—Single-frequency amplitude characteristics of an over-all frequency-modulation radio system.

The deviation from normal transmission, at very small values of h, changes with r in a manner portrayed in Fig. 3, which shows over-all system deviations from normal transmission as a function of r. These were computed from the limiting value of D_1 given in (13) and substituted in (10) in the appendix.

Let us return now to the factors f_d/f_p and $\sin\alpha/2$ in the abscissa. It will be remembered these were made unity for the purposes of the above discussion as well as for establishing a convenient reference for other characteristics in which these quantities are not unity. In practice these quantities are most likely not to be

[6] The values of β and α used in the discussion here are limited to the range 0 to 2π because the numerical values of the equations identically repeat in successive 2π intervals. When β is given any value with respect to α it is supposed that the phase-frequency characteristics of the two paths have such displacement and slope within the transmitted band that $\beta+2\pi n$ coincides with α.

unity. For instance, the deviation index f_d/f_p is inversely proportional to the modulating frequency and has a different value for every frequency in the modulating-frequency band. It is also directly proportional to the maximum carrier-frequency deviation and for any given modulating frequency it has a different value in radio systems having different deviation ratios. A correction may be applied directly to the curves of either Fig. 1 or Fig. 2 so that they will show the amplitude characteristics of a system for any value of deviation index. For example, on a frequency-modulation radio system having a deviation ratio of 5, f_d/f_p would be 5 for the top frequency in the modulating-frequency band.[7] For this frequency, then, a reading on the abscissa scale of Fig. 1 would be increased 14 decibels to obtain the correct ordinate reading of the receiver output. For every value of h the curves would be read 14 decibels to the right.

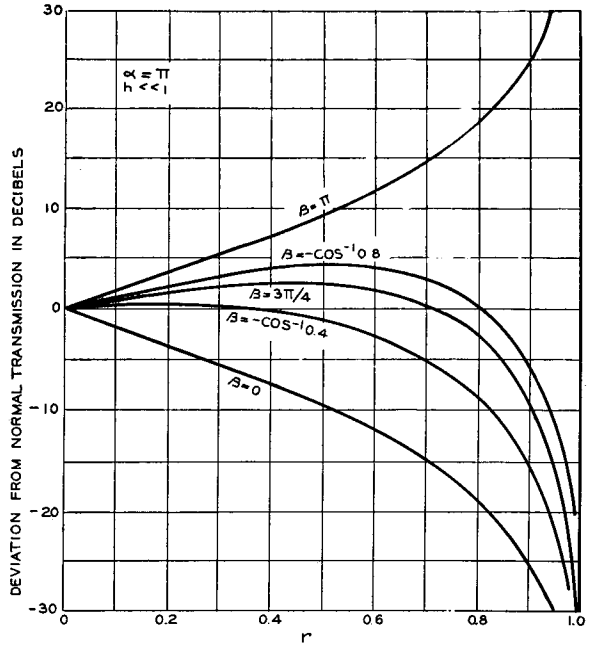

Fig. 3—Over-all-system transmission deviations as a function of path-loss amplitude ratio.

In the general case the curves are read the requisite number of decibels to the right or left of a given abscissa according to whether the corresponding value of f_d/f_p is greater or less than one.

Variations in the relative modulating-frequency phase shift α may be applied directly to the curves of Fig. 1 and Fig. 2 for qualitative purposes only. While the factor $\sin \alpha/2$ in the abscissa may be applied to give shifts in abscissa readings in a manner similar to that for the deviation index, it should be noticed that the shape of the curves is altered when α is varied. This is brought about by the factor $\sin^2 \alpha/2$ in the fundamental term of (10). This factor, as α departs from π, tends to reduce the deviations from normal system transmission from the maximum values shown in Figs. 1 and 2.

[7] At lower modulating frequencies values of f_d/f_p would be proportionately greater.

The points where the various characteristics cross the normal remain fixed, however. This tends to keep the critical region about in the position shown. This is an aid in a qualitative understanding of the effects of variation in α. However, if it is necessary to know the transmission changes accurately for quantitative purposes, it is necessary to plot new sets of curves for different values of α.

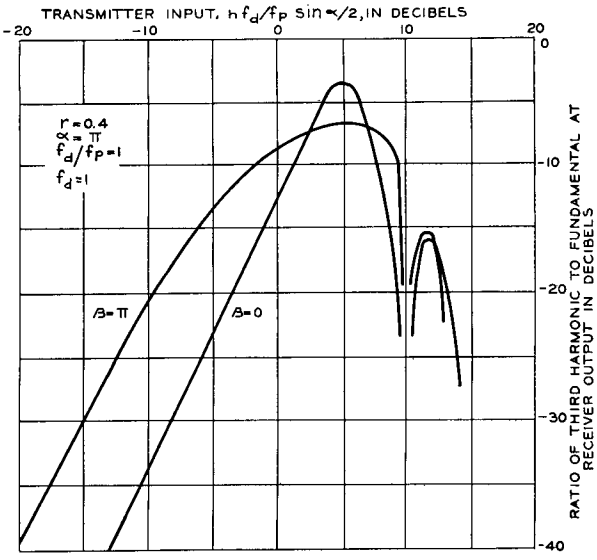

Fig. 4—Single-frequency harmonic output as a function of fundamental input to transmitter.

Before attempting to interpret the curves of Figs. 1 and 2, it may be well first to proceed to a discussion of some of the modulation products attending amplitude characteristics of this type. In Figs. 4 and 5 are shown

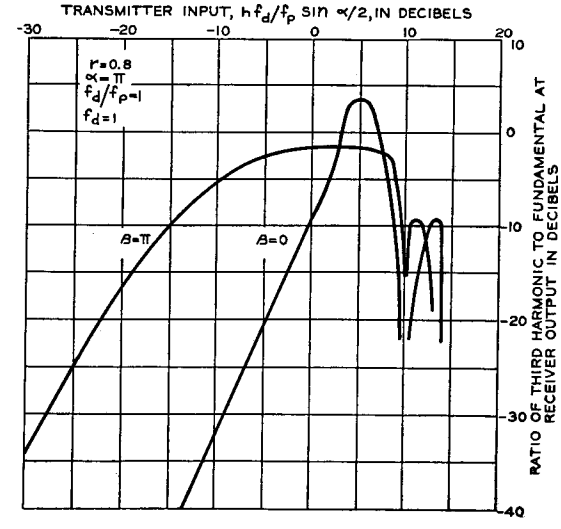

Fig. 5—Single-frequency harmonic output as a function of fundamental input to transmitter.

computed third-harmonic characteristics that go with the amplitude characteristics for $\beta = 0$ and π in Figs. 1 and 2, respectively. These curves are obtained from H, the harmonic-distortion part of (10), as expressed by D in (9) when $m > 1$. The curves represent the ratio of har-

monic to fundamental amplitude at the receiver output as expressed by the ratio D_3 to the fundamental term of (10). The values of D_3 for Figs. 4 and 5 are given in Figs. 10 and 11, respectively, in the appendix. Even-order distortion is not present because as seen from (9) it vanishes when $\beta = n\pi$.

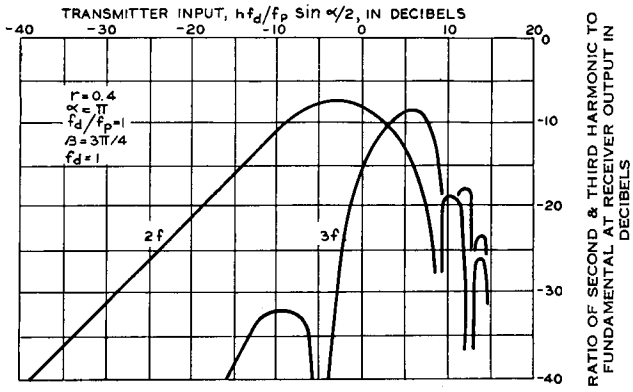

Fig. 6—Single-frequency harmonic output as a function of fundamental input to transmitter.

The co-ordinates in Figs. 4 and 5 are expressed in decibels. As in Figs. 1 and 2, the abscissas are proportional to transmitter input with f_d/f_p and sin $\alpha/2$ both unity. The ordinates are proportional to harmonic output expressed in decibels below the fundamental at the receiver output for unit value of f_d. It is seen that third-harmonic distortion is high, for most uses, in the critical

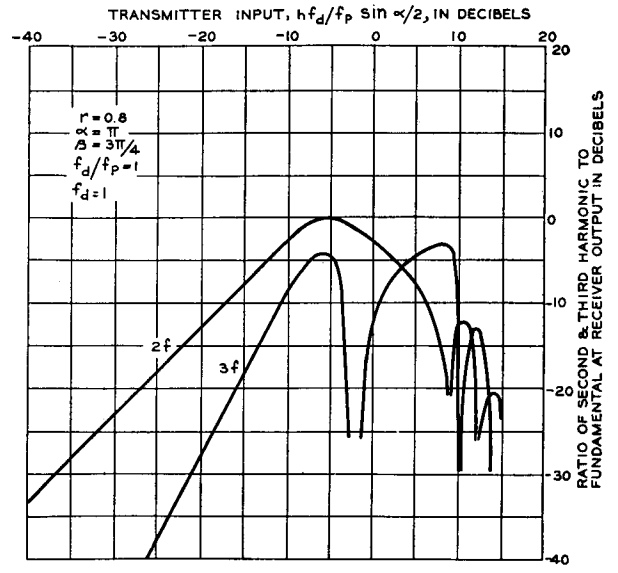

Fig. 7—Single-frequency harmonic output as a function of fundamental input to transmitter.

region. On each side of this region the distortion becomes less. Because D_3 is a multiplier in the harmonic amplitude the sharp dips in the characteristics are at points where D_3 passes through zero. The sign of the harmonic in both curves alternates in successive lobes. At low input amplitudes the harmonic behaves like normal third-order modulation and follows a cube-law characteristic. This is shown analytically in (12) in the appendix.

The harmonic associated with the amplitude characteristic for $\beta = 3\pi/4$ in Figs. 1 and 2 is shown in Figs. 6 and 7. Both $2f$ and $3f$ are present and are plotted to the same scales as $3f$ in the previous set of curves. These curves exhibit the same general properties as those in the previous set except that at small inputs $2f$ decreases half as fast as $3f$.

Higher harmonics than the second and third may be plotted in similar manner. It is sufficient for the present purpose, however, to go no further than these two.

The curves of harmonic output shown here may be considered reference curves. Since the abscissa co-ordinates are identical with those for the fundamental or amplitude characteristics, they are manipulated in the same way for various values of deviation index and sin $\alpha/2$. However, as in the case of the curves for fundamental, the amplitude changes with sin $\alpha/2$ so that harmonic distortion decreases as α departs from π. The points at which the sharp dips occur, however, remain fixed. If

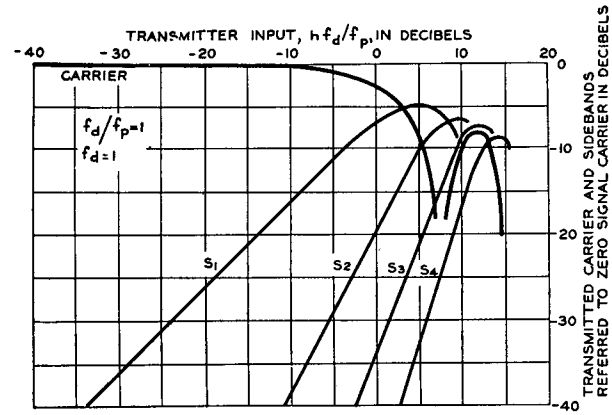

Fig. 8—Carrier and sideband distribution as a function of transmitter input.

accurate information is desired for quantitative purposes, a new set of characteristics must be plotted for the required value of α.

As an additional aid to the understanding of the amplitude nonlinearity depicted by the amplitude and harmonic characteristics of Figs. 1, 2, 3, 4, 5, 6, and 7, a set of curves, Fig. 8, has been arranged to show the relative strength of transmitted carrier and associated sidebands as a function of transmitter input. The co-ordinates are expressed in decibels. As the relation between carrier and sidebands depends only on hf_d/f_p the factor sin $\alpha/2$ is omitted from the abscissa. To determine the wave make-up at the transmitter output associated with any point on the curves in Figs. 1, 2, 3, 4, 5, 6, and 7, it is only necessary to know the associated value of hf_d/f_p and apply it to Fig. 8.

For $\alpha = \pi$ the abscissa zero of Fig. 8 coincides with those of Figs. 1, 2, 3, 4, 5, 6, and 7. A convenient picture of the sideband distribution for a load or harmonic characteristic may be obtained by considering Fig. 8 superimposed on the other figures with abscissa zeros coinciding. When α is other than π the abscissa zero of Fig. 8 is moved to the left of the others the number of decibels

corresponding to sin $\alpha/2$. A study of the characteristics in this manner shows that the critical region in the load characteristics is reached with smallest number of sidebands (about two) when $\alpha=\pi$. As α departs from π the number of sidebands necessary to reach this point increases until when $\alpha=0$ it would take an infinite number of sidebands. Above the critical region the number of sidebands in all cases is large.

Something should be said now concerning alterations of the over-all-system frequency characteristic due to two-path interference. By the original assumption the phase-versus-frequency characteristic in the radio-frequency band varies linearly with frequency making α a linear variable with frequency. Curves which are proportional to the frequency characteristic of an over-all system may be obtained if the fundamental output at the receiver is plotted against α. In Fig. 9 the corresponding deviations from normal system transmission

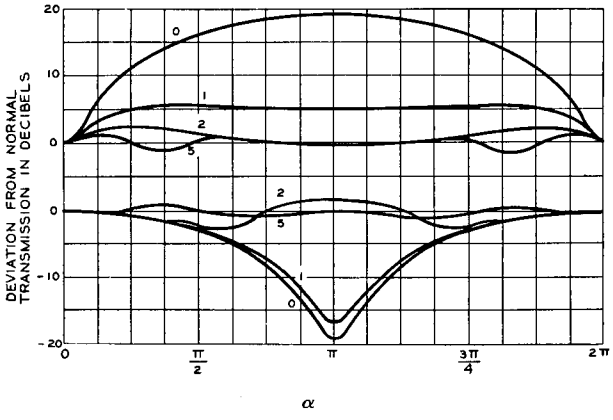

Fig. 9—Frequency distortion; α=linear variable with frequency. The number on each curve is the value of hf_d/f_p used in computation. $\beta=\pi$ in the upper group; $\beta=0$ in the lower group. $r=0.8$.

are plotted against α. These were obtained from (10) and the associated values of D_1 in Fig. 11. The curves are computed for $r=0.8$ and $\beta=0$ and π with the factor hf_d/f_p serving as parameter. The upper group represents $\beta=\pi$ and the lower group $\beta=0$. It is evident that for a given value of deviation index the various characteristics of Fig. 9 will be obtained at different values of h. It should be expected, then, that the frequency characteristic of an over-all system would vary with signal-input level to the transmitter.

Viewing amplitude nonlinearity as a whole, we see from Figs. 1, 2, 3, 4, 5, 6, and 7 that if at any modulating frequency at the transmitter input h were made to vary over a wide range of values, a wide variety of transmission changes would be introduced by the over-all system. It should be remembered in the cases shown that the characteristics are all a function of the product of input amplitude h, deviation index, and sin $\alpha/2$. Because the deviation index is inversely proportional to modulating frequency, the input amplitude necessary to obtain these characteristics is directly proportional to modulating frequency. It results that over a band of modulating frequencies, such as a speech or program band, sudden changes in transmission as shown in Figs. 1 and 2 would occur at different input levels depending on the part of the band in which the assumed conditions of phase and amplitude are realized. The curves for $\beta=0$ could be used to explain what has been termed "volume bursts" in frequency-modulation reception brought about by fading. If the delay between paths is sufficient to satisfy the conditions for these curves at some frequency in the audible band, a passage of the input level upward through the critical point will increase the output amplitude giving a sudden increase in loudness. This sudden change will, of course, be attended by a whole spectrum of harmonics as illustrated. A decrease in level back to the starting point will pass through the critical point in reverse, giving a sudden decrease in loudness. If the change in input level is up and down through this point in a short interval, the effect will be a change corresponding to impulse.

If voice or program material is applied to the transmitter input and amplitude characteristics similar to the above exist, it would be possible to obtain sudden changes in receiver output as the variable amplitude of this material passed haphazardly up and down through critical values of input amplitude. The effect on the listening ear could very well be that which has been described as volume bursts.

Some degree of theoretical corroboration of experimental observations is established in the harmonic characteristics shown. In the operation of actual circuits having multiple-path transmission it has been observed by Crosby[1,2] and others that harmonic distortion appears to be worse at low modulating frequencies than at high. On the other hand, it has been observed by Corrington[3] that the reverse is true. With harmonic characteristics as shown in Figs. 4, 5, 6, and 7, it is apparent that if hf_d/f_p is large enough to reach beyond the region of maximum distortion (critical region), a general reduction in distortion follows a decrease in f_p for a given value of transmitter input (represented by h). This agrees with the second observation. But if the transmitter input is reduced, distortion increases (except for the sharp dips) to a maximum in the critical region and then decreases steadily. If hf_d/f_p is small enough to maintain system operation below the critical region, an increase in distortion follows a reduction in f_p for a given value of transmitter input. This agrees with the first observation. Either observation, then, is true depending on how hard the transmitter is driven.

We have seen from the foregoing what variations in over-all transmission may be expected in frequency-modulation radio systems having single-frequency modulation of the carrier and two-path propagation between transmitter and receiver. Because multiple-path propagation produces amplitude nonlinearity in the system, the principle of superposition cannot be applied to determine what might be expected in over-all performance as the number of paths or the number of frequencies are increased. But what has been shown here may be used

as a gauge to explain in an approximate way some observations that have been made on over-all transmission in frequency-modulation radio circuits.

Appendix

Let e_1 and e_2 represent two received sinusoidal frequency-modulated waves delayed one with respect to the other t_o seconds. Let E_0 and E_1 be the peak amplitudes of the earlier wave and the later wave, respectively. Then at the limiter input

$$e_1 + e_2 = E_0 \sin\left(\omega_c t + h\frac{\omega_d}{\omega_p}\cos\omega_p t\right) + E_1 \sin\left[\omega_c(t+t_0) + h\frac{\omega_d}{\omega_p}\cos\omega_p(t+t_0)\right] \quad (1)$$

where

$\omega_c = 2\pi \times$ carrier frequency
$\omega_p = 2\pi \times$ modulating frequency
$\omega_d = 2\pi \times$ maximum-deviation frequency
$t_0 =$ delay in seconds
$h =$ factor proportional to the amplitude of signal input at the transmitter.

Let $\beta = \omega_c t_0$ and $\alpha = \omega_p t_0$. Then following a procedure identical with that used by Crosby[1] in solving the same case, the desired frequency-modulated wave leaving the limiter is found. The resultant instantaneous frequency is

$$\left.\begin{aligned} f_i &= f_c - hf_d \sin\omega_p t + D, \\ D &= -hf_d 2 \sin\frac{\alpha}{2}\,\frac{r+\cos\theta}{\frac{1}{r}+r+2\cos\theta}\cos\left(\omega_p t + \frac{\alpha}{2}\right) \end{aligned}\right\} \quad (2)$$

where $$r = \frac{E_1}{E_0}$$

and

$$\theta = \beta - l\sin\left(\omega_p t + \frac{\alpha}{2}\right) \qquad l = 2h\frac{f_d}{f_p}\sin\frac{\alpha}{2}.$$

If the discriminator and detector circuits are assumed to be distortionless, the detector output will be proportional to f_i, so we may confine our analysis to f_i. Accordingly, the first term of (2) is a constant giving rise to direct current. The second term is the fundamental component of the modulating frequency having a peak amplitude of hf_d. The third term is the distortion produced by two-path interference and is of immediate interest. By application of Fourier analysis to that part of D involving $\cos\theta$ an expansion into a series involving r and $\cos\theta$ may be obtained. The result is

$$\frac{r+\cos\theta}{\frac{1}{r}+r+2\cos\theta} = -\sum_{n=1}^{\infty}(-r)^n\cos n\theta, \quad r<1 \quad (3)$$

making use of the value of θ given in (2)

$$\left.\begin{aligned} \cos n\theta &= \cos\left[n\beta - nl\sin\left(\omega_p t + \frac{\alpha}{2}\right)\right] \\ &= \cos(n\beta)\cos\left[nl\sin\left(\omega_p t + \frac{\alpha}{2}\right)\right] \\ &\quad + \sin(n\beta)\sin\left[nl\sin\left(\omega_p t + \frac{\alpha}{2}\right)\right] \end{aligned}\right\}. \quad (4)$$

This may be expanded so that

$$\cos n\theta = \sum_{m=-\infty}^{\infty} J_m(nl)\left[\cos(n\beta)\cos m\left(\omega_p t + \frac{\alpha}{2}\right) + \sin(n\beta)\sin m\left(\omega_p t + \frac{\alpha}{2}\right)\right] \quad (5)$$

where J_m is an mth order Bessel function of the first kind. On the substitution of (5) into (3) we have

$$\frac{r+\cos\theta}{\frac{1}{r}+r+2\cos\theta} = -\sum_{n=1}^{\infty}\sum_{m=-\infty}^{\infty}(-r)^n J_m(nl)\left[\cos(n\beta)\cos m\left(\omega_p t + \frac{\alpha}{2}\right) + \sin(n\beta)\sin m\left(\omega_p t + \frac{\alpha}{2}\right)\right], \quad r<1. \quad (6)$$

Then substituting (6) in the expression for D in (2) the complete distortion component is obtained and

$$\begin{aligned} D = hf_d \sin\left(\frac{\alpha}{2}\right)\sum_{n=1}^{\infty}\sum_{m=-\infty}^{\infty}(-r)^n J_m(nl) \times \Bigg\{ &\cos(n\beta)\left[\cos(m+1)\left(\omega_p t + \frac{\alpha}{2}\right) + \cos(m-1)\left(\omega_p t + \frac{\alpha}{2}\right)\right] \\ &+ \sin(n\beta)\left[\sin(m+1)\left(\omega_p t + \frac{\alpha}{2}\right) + \sin(m-1)\left(\omega_p t + \frac{\alpha}{2}\right)\right]\Bigg\}. \end{aligned} \quad (7)$$

It will be noticed that in the sine and cosine terms as m progresses from 1 to $\pm\infty$ like frequencies appear at m and $m\pm 2$. Thus when $m=0$ the fundamental of the modulating frequency appears. When $m=\pm 2$ the fundamental and third harmonic appear. When $m=\pm 4$ the third and fifth harmonics appear, etc. Due to this overlapping the coefficients of the sine and cosine terms include the sum of two Bessel functions instead of one. By means of the recurrence formula for Bessel functions of the first kind these may be resolved into one. When this resolution has been made the final expression for the distortion becomes

$$D = hf_d \sin\left(\frac{\alpha}{2}\right) \sum_{n=1}^{\infty} \sum_{m=-\infty}^{\infty} (-r)^n \frac{2m}{nl} J_m(nl) \times \left[\cos(n\beta) \cos m\left(\omega_p t + \frac{\alpha}{2}\right) + \sin(n\beta) \sin m\left(\omega_p t + \frac{\alpha}{2}\right)\right], \quad r < 1. \quad (8)$$

When m is odd, the + and − values of m cause the sine terms to cancel and the cosine terms to add. Conversely when m is even the + and − values of m cause the cosine terms to cancel and the sine terms to add, so that odd-order distortion is represented by the cosine terms and even-order distortion by the sine terms. Some simplification in notation in (8) may be achieved by allowing m to designate the order of the frequency. Then

$$\left.\begin{aligned} D &= hf_d \sin\left(\frac{\alpha}{2}\right) D_m \cdot \cos\left[m\left(\omega_p t + \frac{\alpha}{2}\right) - \frac{\pi}{2} \frac{1 + (-1)^m}{2}\right], \\ D_m &= \frac{4m}{l} \sum_{n=1}^{\infty} (-r)^n \frac{J_m(nl)}{n} \cdot \cos\left[n\beta - \frac{\pi}{2} \frac{1 + (-1)^m}{2}\right]. \end{aligned}\right\} \quad (9)$$

Either (8) or (9) shows the composition of the distortion component at the receiver output in terms of the fundamental and odd and even harmonics of the modulating frequency.

Several interesting points may be gathered from these equations. $D = 0$ when the attenuation ratio of the two paths is 0, which is to be expected. For in such a case one path has no transmission and no interference exists. $D = 0$ when the relative phase shift of the modulating frequency is $\alpha = 2\pi n$, because the series summation is finite when $r < 1$ and $\sin(\alpha/2) = \sin \pi n = 0$. As α is a function of relative delay it is possible to infer from this that distortion is small when the relative delay between paths is small. $D = 0$ when l is very large because

$$J_m(nl) \cong \left(\frac{2}{\pi nl}\right)^{1/2} \cos\left(nl - \frac{\pi}{4} - \frac{m\pi}{2}\right) \doteq 0, \quad l \doteq \infty$$

and when substituted in D_m the series summation likewise approaches zero when $r < 1$. But l is large when hf_d/f_p is large. We may infer from this that a system having high input amplitude or high deviation index or both will have small distortion.

These points may be seen more clearly in the curves of D_m versus l shown in Figs. 10 and 11. Here D_1, D_2, and D_3 (fundamental, second, and third harmonic) are plotted against l for three values of relative carrier phase shift between paths, $\beta = 0$, $3\pi/4$, and π; and amplitude ratios of the path losses of $r = 0.4$ and 0.8. Each point on the curves involves a summation of the series for D_1, D_2, and D_3. D_2 is absent when $\beta = 0$ and π because even-order modulation vanishes when $\beta = \pi n$.

The curves for D_1 are shown on the lower half of Figs. 10 and 11. These exhibit the same general properties as the curves for D_2 and D_3. However, D_1, being a component of fundamental in the distortion term D, must be combined with the regular fundamental term in (2) to give the complete output fundamental component. Such a combination alters (2) so that

$$\left.\begin{aligned} f_i &= f_c - hf_d \sqrt{1 + D_1(2 + D_1) \sin^2\frac{\alpha}{2}} \cdot \sin(\omega_p t - \phi) + H, \\ \phi &= \tan^{-1} \frac{D_1 \sin \alpha}{2\left(1 + D_1^2 \sin^2\frac{\alpha}{2}\right)} \\ H &= D \quad \text{for} \quad m > 1 \end{aligned}\right\} \quad (10)$$

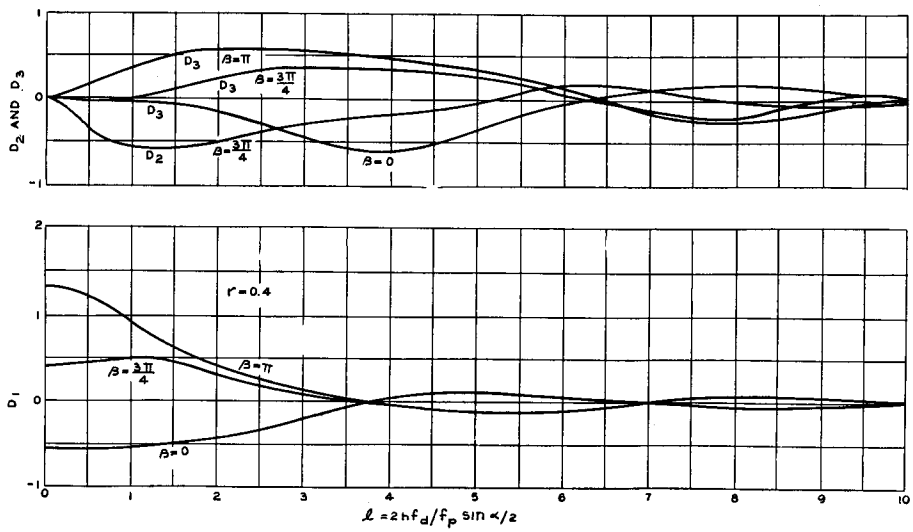

Fig. 10—D_m as a function of l for the indicated values of r, β, and m.

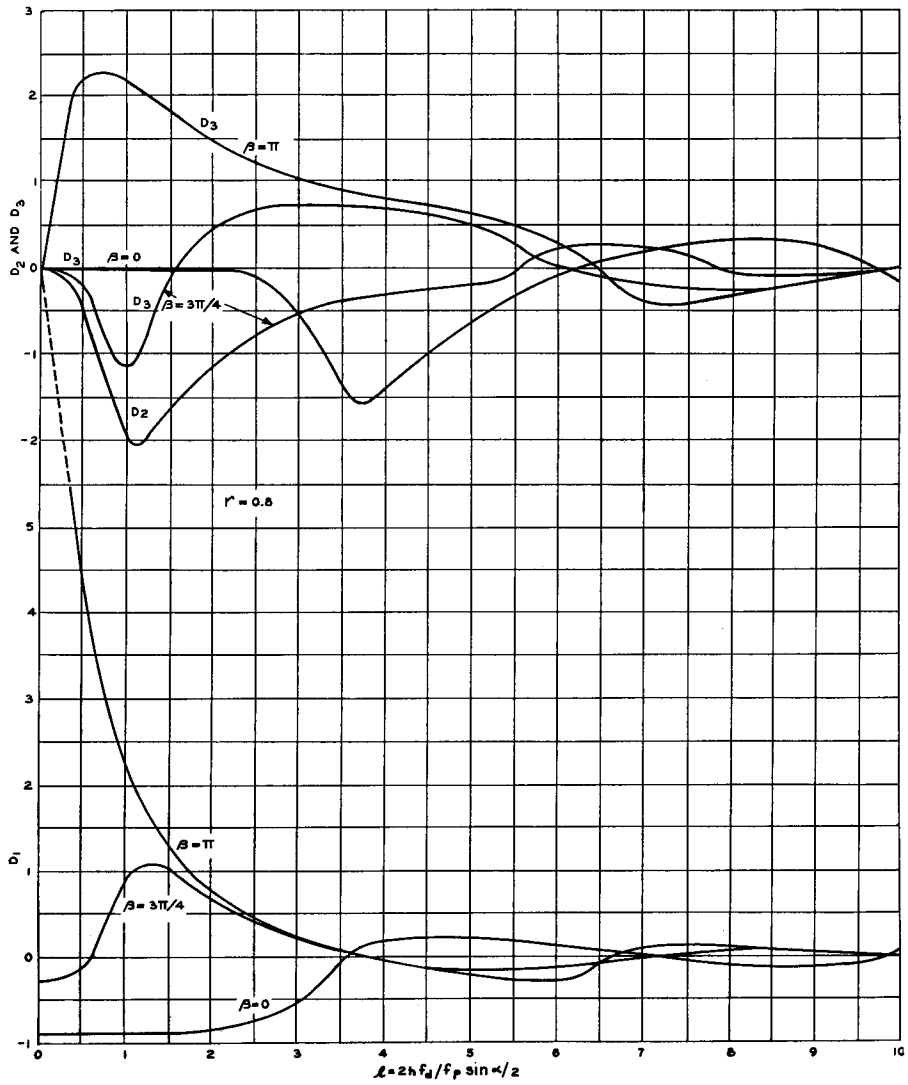

Fig. 11—D_m as a function of l for the indicated values of r, β, and m.

The coefficient of the complete fundamental term now increases and decreases with positive and negative values of D_1 above and below its normal value for zero distortion. H is used here to represent harmonic distortion in D and is equal to D in (9) with the fundamental term removed.

It is sometimes of interest to know how the distortion varies at low-signal amplitudes such as shown in Figs. 2 and 3 in the text. This may be obtained from (9) for very small values of h by replacing $J_m(nl)$ by the first term of the Bessel's series expansion. The result is

$$D_m \sim 2\frac{\left(\frac{l}{2}\right)^{m-1}}{(m-1)!}\sum_{n=1}^{\infty}(-r)^n(n)^{m-1}\cdot\cos\left[n\beta-\frac{\pi}{2}\frac{1+(-1)^m}{2}\right] \quad (11)$$

letting

$$\psi=\sum_{n=1}^{\infty}(-r)^n(n)^{m-1}\cos\left[n\beta-\frac{\pi}{2}\frac{1+(-1)^m}{2}\right]$$

we have

$$D\sim h^m\frac{2f_d\sin^m\left(\frac{\alpha}{2}\right)\left(\frac{f_d}{f_p}\right)^{m-1}}{(m-1)!}\psi\cdot\cos\left[m\left(\omega_p t+\frac{\alpha}{2}\right)-\frac{\pi}{2}\frac{1+(-1)^m}{2}\right]. \quad (12)$$

From (12) the amplitude of the mth-order harmonic varies as h^m which is proportional to mth power of the input amplitude.

A summation of (11) is most readily made for specific values of m. For present purposes, where fundamental, second, and third harmonic only are considered, it is sufficient to let $m=1, 2, 3$ to obtain the necessary limiting values of D_m as $l\doteq 0$. The resulting values of D_m are

$$D_1 \sim -2r \frac{r+\cos\beta}{1+r^2+2r\cos\beta}$$
$$D_2 \sim -\frac{lr(1-r^2)\sin\beta}{(1+r^2+2r\cos\beta)^2} \qquad (13)$$
$$D_3 \sim -\left(\frac{l}{2}\right)^2 r(1-r^2)\frac{2r(1+\sin^2\beta)+(1+r^2)\cos\beta}{(1+r^2+2r\cos\beta)^3}.$$

As h is a factor in l, these expressions are also valid for $h \doteq 0$, making them useful in computing load and harmonic characteristics for small values of h.

ACKNOWLEDGMENT

The author wishes to express appreciation to his colleagues for many helpful suggestions and constructive criticism generously offered during the preparation of this paper.

Discussion on

"Concerning Hallén's Integral Equation for Cylindrical Antennas"*

S. A. SCHELKUNOFF

Ronold King:[1] The integral equation of Hallén in the form

$$\int_{-l}^{l} \frac{I(\xi)e^{-i\beta r}}{r}\,d\xi = A\cos\beta z + B\sin\beta|z| \qquad (1)$$

applies to a perfectly conducting cylinder of sufficiently small cross section so that the axial distribution of current may be treated as independent of the transverse distribution. The evaluation of the constants A and B is carried out subject to the idealized boundary condition $I(l)=0$ and in terms of an idealized "slice" generator which is exactly like the rest of the conductor except for a discontinuity in scalar potential at the center, $z=0$. This discontinuity in scalar potential is not in the form of a dielectric gap between the halves of the antenna placed end to end, but is assumed to exist at the center of a continuous conductor. The degree in which these conditions may be approximated in practice has been discussed in some detail in another place[2] where it is pointed out and experimentally verified that an antenna which is center driven from a two-wire line does not approximate well the assumed driving conditions. This is due both to the gap and to the more or less adjacent end surfaces of the halves of the antenna, the two having somewhat compensating effects. Neither an infinitesimal gap nor a wide gap (as suggested by Dr. Schelkunoff) is a satisfactory experimental equivalent of the "slice" generator assumed in the theory. While an exact physical counterpart of a "slice" generator is unavailable, the best approximation would seem to be achieved in a base-driven antenna over a sufficiently large conducting plane. If the antenna is made the continuation above the plane of the inner conductor of a coaxial line, with inner and outer conductors nearly equal in size, a significant rotationally symmetrical driving field is maintained only over a very short distance along the base of the antenna in a manner resembling a "slice" generator with skin effect.

Dr. Schelkunoff's comparison of the expansions due to Hallén and to Gray may be supplemented by a third method of expanding the integral on the left in (1) that appears to be very much superior to both of those discussed. Let the three methods be considered in turn. The Hallén method proceeds from

$$\int_{-l}^{l} \frac{I(\xi)e^{-i\beta r}}{r}\,d\xi = I(z)\int_{-l}^{l}\frac{d\xi}{r} + \int_{-l}^{l}\frac{I(\xi)e^{-i\beta r}-I(z)}{r}\,d\xi, \qquad (2)$$

which is equivalent to assuming as a first approximation a *uniform* current along the entire antenna equal to the current $I(z)$ at the point of calculation, and, in addition, to neglecting retardation. It leads to the simple expansion parameter $\Omega = 2ln(2l/a)$ that is independent of βl.

The assertion by Dr. Schelkunoff that "Hallén's first approximation involves a tacit assumption that the antenna is short compared with the wavelength" appears to rest solely on the argument (p. 875) that the first term on the right in (2) is "a better approximation to the integral on the left if βr is small; that is, if $2\pi l/\lambda$ is small." No proof is given of this statement and it certainly is *not correct* in general. This is easily reasoned out qualitatively, but is perhaps most quickly and convincingly verified by computing a simple numerical case. For a short antenna which satisfies the condition $\beta l < 0.5$ the distribution of current is very well approximated by $I(\xi) = I(0)(1-[|\xi|/l])$. If this is substituted in the integral on the left in (2) the result is $I(0)[\Omega - 2 - j\beta l]$. The first integral on the right gives $I(0)\Omega$. The

* PROC. I.R.E., vol. 33, pp. 872–878; December, 1945.

[1] Cruft Laboratory, Harvard University, Cambridge, Mass.

[2] R. King and D. D. King, "Microwave impedance measurements with application to antennas, II," *Jour. Appl. Phys.*, vol. 16, pp. 445–453; August, 1945.