The Phase Problem

R. E. Burge; M. A. Fiddy; A. H. Greenaway; G. Ross

Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences, Vol. 350, No. 1661. (Aug. 20, 1976), pp. 191-212.

Stable URL:
http://links.jstor.org/sici?sici=0080-4630%2819760820%29350%3A1661%3C191%3ATPP%3E2.0.CO%3B2-A

Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences is currently published by The Royal Society.

Proc. R. Soc. Lond. A. **350**, 191–212 (1976)
Printed in Great Britain

The phase problem

By R. E. Burge, M. A. Fiddy, A. H. Greenaway† and G. Ross
Department of Physics, Queen Elizabeth College, Campden Hill Road, London W8 7AH, England

(*Communicated by W. C. Price, F.R.S.* – *Received* 20 *August* 1975 – *Revised* 18 *February* 1976)

The paper discusses the use of the theory of entire functions for solving the phase problem. In all practical cases only three forms of logarithmic Hilbert transform could possibly be required. The paper defines them and analyses their applicability. A generating form is also put forward for cases of possible theoretical interest. The uniqueness of the phase obtained from a logarithmic Hilbert transform is investigated and the difficulties due to the presence of zeros in the complex plane are discussed. Methods are put forward for both the removal of the zeros and, when this is not possible, for locating them in order to include their effect. The paper analyses known experimental methods for phase determination from the point of view of the theory presented and highlights their unique character.

1. Introduction

Detecting visible light and radiation – electromagnetic or de Broglie – of higher frequencies provides data related to the energy density of the radiative field. The field however, in a quasimonochromatic situation, is characterized by two parameters: the modulus and phase, i.e. by a complex function. The intensity is proportional to the square of the modulus and phase information is lost on recording.

Let us consider, for example, a scattering/image formation experiment; we shall designate the field in the object (primary) space by $\mathscr{E}(p)$ and the scattered wave in the Fraunhofer space by $E_s(s)$. If an image is formed, the field in the image space is $E_i(p')$. The complex amplitudes $\mathscr{E}(p)$, $E_s(s)$ and $E_i(p')$ are related by Fourier transform relations and hence the determination of any one of these functions provides the others. In our scattering experiments we determine only $|E_s(s)|$ and/or $|E_i(p')|$.

In general, determination of $|E_i(p')|$ provides only geometrical information concerning the object while knowledge of $|E_s(s)|$ permits determination only of the statistics of $\mathscr{E}(p)$. In principle, this may provide information on the statistical properties of the object itself.

The object of this paper is the determination of the complex object wave. This requires the solution to the phase problem, i.e. the determination of $E_s(s)$ or $E_i(p')$,

† Present address: Technical Physical Laboratory, State University at Groningen, Nijenbourgh 18, Groningen 8002, The Netherlands.

from the measured $|E_s(s)|$ and/or $|E_i(p')|$. The problem of the relation between the object and the object wave is not considered here.

The phase problem occurs not only in scattering phenomena as introduced here, but also in image analysis, radio astronomy, coherence theory, and other related areas. Our analysis can be extended to any phase retrieval problem when a Fourier space can be defined. For generality we shall assume, un less specified otherwise, that the measurements are made either in the Fraunhofer space *or* the image space. The function in the space of measurement is denoted by $F(x)$ and the function in the conjugate space is $f(t)$.

The principal aims are to establish the conditions under which a solution to the phase problem can be found, and to seek methods of its solution. Because of the Fourier relation between the complex amplitudes as well as the finite extent of the intervals in which these amplitudes are physically defined, we have sought a solution using the theory of entire functions. This approach provides an integral relation between the real and imaginary parts of a function and the Hilbert transform appeared the natural one to consider. Its usual formulation had to be modified and the convergence and uniqueness of the solutions have been examined. The recovery of phase as described here allows arbitrary changes of origin of the coordinate systems in both the primary and Fraunhofer spaces (since a modulus is invariant with respect to such shifts in its Fourier space) and the multiplication of the field in any space by an arbitrary but unimodular constant.

We shall restrict our consideration to one dimensional situations. We shall also consider quasi-monochromatic fields only and exclude from our discussion inelastic effects of any sort.

2. Mathematical background

To formulate the Hilbert transforms in suitable forms, we consider the convergence properties of the appropriate integrals by continuing the function $F(x)$ into the complex plane and examining its behaviour, especially at infinity.

The complete description of this behaviour is provided by the theory of entire functions (see, for example, Boas (1954) and Cartwright (1955)), and in particular by the properties of the Phragmén–Lindelöf function.

The following relation shows the significance of the two assumptions made, namely that a Fourier relation exists in scattering experiments and that objects are of finite extent

Theorem 1. If a function $F(z)$ is defined by

$$F(z) = \int_a^b f(t) \exp(izt)\, dt, \tag{2.1}$$

where $z = x + iy$, $f(t)$ integrable in the interval (a, b), $|a| \leqslant b < \infty$, and $f(t)$ does not vanish almost everywhere in any neighbourhood of a and b, then $F(z)$ is an entire

function of order 1 and type b (and not smaller type)† and refers to all functions of physical interest.

In general, the growth of $|F(z)|$ is anisotropic. A descriptor of this direction-dependent growth is the Phragmén – Lindelöf function (or indicator function), $h(\theta)$. For a function of unit order,

$$h(\theta) = \limsup_{r\to\infty} \frac{\ln |F(re^{i\theta})|}{r}. \tag{2.2}$$

The indicator function may be identified with a supporting function.‡ As such, its value for any direction θ may be found from the indicator diagram.§ For functions defined by equation (2.1), the indicator diagram is the line segment $(-ib, -ia)$. The indicator function is the most positive projection of this indicator diagram onto the ray $\arg z = \theta$, ($h(\theta)$ is negative if the projection lies completely on $\arg z = \theta + \pi$). Thus if, for example, $a = -b$, the line segment is $(-ia, ia)$ and we have, for the projection onto the ray $\arg z = \theta$,

$$h(\theta) = a\,|\sin\theta|. \tag{2.3}$$

For $b > |a|$, the maximum value of $|h(\theta)|$ is b and lies in the direction $\theta = -\frac{1}{2}\pi$. Clearly, for functions defined by (2.1), $h(\theta) = 0$ for $\theta = 0$ or π which means that $F(x)$ behaves, as $|x| \to \infty$, as a finite power of x (and cannot behave exponentially on the real axis). Since $F(x) \in L^2$, $F(x)$ tends to zero as $|x| \to \infty$. Thus, from the above, it follows that, as $|x| \to \infty$, $F(x) \to 0$, as a finite power.

There are three possible situations to be considered for the positions of the indicator diagram relative to the origin.

THEOREM 2. For functions defined by equation (2.1), $h(\theta)$ has the following properties:

(i) If the origin is inside the line segment $(-ib, -ia)$, i.e. $a < 0$, then $h(\theta) \geqslant 0$ for all θ, the equality holding only along the real axis.

(ii) If the origin is outside the line segment $(-ib, -ia)$, i.e. $a > 0$, then $h(\theta)$ is negative only for direction $\arg z = \theta$ projected into the half plane not containing the segment, i.e. the upper half plane.

(iii) If the origin coincides with a, then $h(\theta)$ is zero for directions not containing the segment, i.e. the upper half plane.

† The order, ρ, and type σ, are defined by

$$\rho = \limsup_{r\to\infty} \frac{\ln \ln [\max |F(r)|]}{\ln r}, \quad \sigma = \limsup_{r\to\infty} \frac{\ln [\max |F(r)|]}{r^\rho},$$

where the supremum means the least upper bound, and $z = re^{i\theta}$. The importance of the two parameters in this application is that they represent a limit of the maximum growth rate of $|F(z)|$ as $|z| \to \infty$.

‡ The support function associated with some $\Omega(\eta)$ defines the domain in which $\Omega(\eta)$ does not vanish.

§ The indicator diagram is a non-empty, bounded, closed, convex set of points the projection of which onto the direction $\arg z = \theta$ is a support function.

So far, we have discussed the maximum rate of growth of $F(z)$. However, of more interest in this application, is the actual rate of growth of $F(z)$. In certain circumstances, we are able to use $h(\theta)$ to describe the actual rate of growth by dropping the supremum from equation (2.2), i.e.

$$h(\theta) = \lim_{r\to\infty} \frac{\ln|F(r\,\mathrm{e}^{\mathrm{i}\theta})|}{r}. \tag{2.4}$$

In order to write (2.4) for zero free angles, two conditions must be met:

(i) $h(0)+h(\pi) = 0$, and (ii) $F(z)$ must have a finite density of zeros.

Since in our case, $h(0) = h(\pi) = 0$, the first condition is fulfilled; if $n(r)$ is the number of zeros of $F(z)$, $|z| \leqslant r$, then the second condition is verified too, provided that $n(r)/r$ is finite for arbitrarily large r. This is known to be true, since Titchmarsh (1925) has established that

$$\lim_{r\to\infty} \frac{n(r)}{r} \sim \frac{(b-a)}{\pi}. \tag{2.5}$$

The influence of the limits a and b on the actual growth of $|F(z)|$ in the complex plane can now be described.

When $a \geqslant 0$, $F(x)$ is referred to as a causal transform† and Titchmarsh's theorem is applicable (Nussenzveig 1972).

THEOREM 3. (Titchmarsh's Theorem.) If a function $F(z)$ fulfils one of the following conditions, it fulfils all of them.

(i) $F(z)$ is for almost all x, the limit as $y \to 0^+$ of an analytic function $F(z)$ regular for $y > 0$ and $\in L^2$ over any line in the u.h.p. which is parallel to the real axis, i.e.

$$\int_{-\infty}^{+\infty} |F(z)|^2\,\mathrm{d}x < c \quad (y > 0).$$

(ii) The inverse Fourier transform, $f(t)$, of $F(x)$ vanishes for $t < 0$ (i.e. $0 \leqslant a < b$ in theorem 1).

(iii) Real $F(x)$ and imaginary $F(x)$ verify the formula

$$\operatorname{Re} F(x') = \frac{1}{\pi} P\int_{-\infty}^{\infty} \frac{\operatorname{Im} F(x)\,\mathrm{d}x}{x-x'}. \tag{2.6}$$

(iv) Real $F(x)$ and imaginary $F(x)$ verify the formula

$$\operatorname{Im} F(x') = -\frac{1}{\pi} P\int_{-\infty}^{\infty} \frac{\operatorname{Re} F(x)\,dx}{x-x'}. \tag{2.7}$$

† In temporal processes the requirement $f(t) = 0$ for $t \leqslant a$, for $a \geqslant 0$, is referred to as the causality condition and may be associated with the statement that 'the effect cannot precede the cause'. Here, where discussion is concerned with functions of spatial rather than temporal variables such a concept would seem misplaced. However, we retain the terminology 'causal transform' to mean a function whose Fourier transform vanishes for negative values of its argument.

The relations (2.6) and (2.7) are known as Hilbert transforms (or dispersion relations): the P denotes that the Cauchy principal value is to be taken. Thus when $0 \leqslant a < b$ we have an integral relation between the real and imaginary parts of $F(x)$. For the particular case when $a = 0$ we shall designate that $F(z)$ be an $\mathcal{O}$-transform (see § 3).

3. The logarithmic Hilbert transform (l.H.t.)

We consider here the direct determination of $F(x)$ from measured $|F(x)|$ by associating the modulus and phase with the real and imaginary parts of a causal transform. We write

$$F(z) = |F(z)| \exp(\mathrm{i}\phi(z)),$$

$$\ln F(z) = \ln|F(z)| + \mathrm{i}(\phi(z) \pm 2n\pi),$$

where n is an integer specifying the Riemann surface in which the (multiple-valued) logarithmic function is defined. Thus the real part of $\ln F(z)$ is determined solely by the modulus and if theorem 3 may be applied to this function we have, in principle, the desired relation between the measured modulus and the required phase.

Unfortunately, in practice, the situation is not as simple as that. It is well known (Titchmarsh 1939) that $\ln F(z)$ has the same region of analyticity as $F(z)$ except at the points where $F(z) = 0$. However the function $\ln F(z)$ is not, in general, a causal transform. This is due to the fact that $F(x) \in L^2$ and thus tends to zero as $|x| \to \infty$. It follows that $\ln F(x)$ diverges to $-\infty$ and hence cannot be square integrable, its Fourier transform cannot be defined, and thus cannot be shown to satisfy theorem 3 (ii). While some authors (see, for example, Bates 1969; Goedecke 1975; Saxton 1975) are aware of this fact, other authors have overlooked or ignored it (see, for example, Wolf 1962; Roman & Marathay 1963; King 1975). This convergence problem has been resolved for a particular case, by Peřina (1971) and considered in an intuitive way for another particular situation by Page (1955). We adopt a more general approach to the problem by either creating a causal transform containing $\ln F(x)$ (§ 3 (*a*)), or modifying $F(x)$ (§§ 3 (*b*) and (*c*)) prior to utilizing the logarithm of its modulus thus achieving a causal logarithmic transform.

The analysis has been suggested by Toll (1956) (see also Hilgevoord 1960) but does not seem to have been previously applied to the Hilbert transform of logarithmic functions. The formulation of the logarithmic Hilbert transform (l.H.t.) for various behaviours of $F(z)$ is considered below. We assume in this paragraph that $F(z) \neq 0$ in a half plane which is consistent with writing (2.4). This will be regarded as the simplest case of a 'reference transform' which will be introduced in § 4 (*a*). The case in which $F(z)$ has zeros in this half plane will be examined in § 4.

(*a*) *The restricted logarithmic Hilbert transform* (*r.H.t.*)

One way to achieve a function having a logarithm which is square integrable is to add a finite constant A, obtaining $F^1(x) = A + F(x)$ such that $\ln F^1(x) \in L^2$. This is satisfied when A is normalized to unity.

We require regularity in a half plane of the function $F^1(x)$ and this implies that $F^1(x)$ is also a causal transform. This is consistent with the indicator diagram approach, as the indicator diagram of a constant is a point at the origin. If $F(x)$ is a causal transform then the indicator diagram of $F(z)$ lies to one side of the origin. The indicator diagram of the sum of these two functions is the smallest convex set containing the union of the separate indicator diagrams, i.e. the line segment $(-ib, 0)$ and so $F^1(x)$ satisfies theorem 3 (iii) and is also an $\mathcal{O}$-transform.

It has been shown (Burge, Fiddy, Greenaway & Ross 1974) that if

$$F^1(x) = |F^1(x)| \exp(i\alpha(x)),$$

then, provided that $F^1(z) \neq 0$ for $y > 0$, the real and imaginary parts of the function $\ln F^1(x)$ are related by Hilbert transform, i.e.

$$\alpha(x') = -\frac{1}{\pi} P \int_{-\infty}^{\infty} \frac{\ln |F^1(x)| \, dx}{x - x'}. \tag{3.1}$$

The phase $\alpha(x')$ is easily related to the phase required $\phi(x')$.

Saxton (1975) and Bates (1969) state that an equation of the form (2.7) in terms of $\ln F(x)$ may give a good approximation to the phase in practice but this could at best be true only if the allowable variation of $|F(x)|$ is restricted. The derivation of (2.7) by Peřina (1971) for $\ln |F(x)|$ requires both real $F(x)$ and imaginary $F(x) \neq 0$, i.e. the addition of a constant.

(*b*) *The modified logarithmic Hilbert transform* (*m.H.t.*)

When $a = 0$ in a half plane (in which the Hilbert transform is evaluated by contour integration) $h(\theta) = 0$ as $r \to \infty$. Hence from (2.4), if $F(x)$ is an $\mathcal{O}$-transform, and if its zeros have a density,

$$\lim_{r \to \infty} \frac{\ln |F(r e^{i\theta})|}{r} = 0, \tag{3.2}$$

then the function $\ln F(z)/z$ is the desired modified function containing $\ln F(z)$.

We write a modified Hilbert transform (m.H.t.) for the function $\ln F(z)$ by taking the real and imaginary parts, on the real axis, of the following integral

$$\int_C \frac{\ln F(z) \, dz}{z(z - x')},$$

where C is a closed contour consisting of the real axis and a semicircle of infinite radius.

Since

$$\frac{\ln F(x)}{x} = \frac{\ln |F(x)|}{x} + \frac{i\phi(x)}{x}, \tag{3.3}$$

the first term on the right hand side of (3.3) converges. However, the convergence of the second term must be investigated. If the phase ϕ is bounded for all z in the

u.h.p. then the second term is clearly zero in the limit as $r \to \infty$ in the u.h.p. If $\mathrm{Re}\, F(x) \neq 0$, it follows that $\mathrm{Re}\, F(z) \neq 0$ in the u.h.p. and the phase will satisfy the restriction $|\phi(z)| = |\arctan(\mathrm{Im}\, F(z)/\mathrm{Re}\, F(z))| < \frac{1}{2}\pi$ for all z, thus ensuring the convergence of (3.3). In addition it implies $F(z) \neq 0$, i.e. no zeros in the u.h.p., a condition which we have already said is assumed. Alternatively the assumption that the phase is bounded is equivalent to considering only the principal part of $\ln F(z)$; this is necessary since we use a closed contour to evaluate the Hilbert integral. Applying the calculus of residues we have

$$\ln F(x') = -\frac{\mathrm{i}x'}{\pi} P \int_{-\infty}^{\infty} \frac{\ln F(x)\,\mathrm{d}x}{x(x-x')} + \ln F(0); \tag{3.4}$$

P again denotes the Cauchy principal value. Taking real and imaginary parts of equation (3.4) gives the dispersion relations

$$\ln |F(x')| = \frac{x'}{\pi} P \int_{-\infty}^{+\infty} \frac{\phi(x)\,\mathrm{d}x}{x(x-x')} + \ln |F(0)|, \tag{3.5}$$

and

$$\phi(x') = -\frac{x'}{\pi} P \int_{-\infty}^{\infty} \frac{\ln |F(x)|\,\mathrm{d}x}{x(x-x')} + \phi(0). \tag{3.6}$$

These relations are correct if $F(z)$ is assumed to behave as a finite power of r as $r \to \infty$ in the u.h.p. since then $r^{-1} \ln |F(r)| \to 0$ as $r \to \infty$ in the u.h.p.; this is always the case when $F(x)$ is an $\mathcal{O}$-transform.

The m.H.t. (equation (3.5)) has been mentioned by Page (1955). He does not investigate the properties of the function in the u.h.p., and his conclusion that the contribution from the integral along the infinite semi-circle at infinity vanishes, is intuitive and without proof.

For all entire functions bounded on the real axis, $h(0) = h(\pi) = 0$ and so the integral in (3.6) will converge whatever the behaviour of $\ln |F(r)|$ in the u.h.p. Since the modulus on the real axis is invariant under a change of origin in the Fourier space, a change in the values of a, b is possible such that $a = 0$ (and $b \to b + a$). This ensures that $h(\theta) = 0$ in a half plane and the solution of (3.6) has the interpretation required, i.e. it is a phase. However it is the phase of an $\mathcal{O}$-transform, which is precisely the original function, phase shifted by an amount which is exactly that required to make $a = 0$. Thus (3.6) may *always* be used for phase recovery in practice, the resulting function being an $\mathcal{O}$-transform irrespective of the original values of a, b.

From a rigorous theoretical point of view the m.H.t. can *only* be applied to regular functions in a half-plane or entire functions which can be made regular in a half-plane by a shift of origin in the Fourier space, i.e. to functions of order 1; this includes all physically realizable functions.

(c) *The general logarithmic Hilbert transform (g.H.t.)*

We consider the case $a \neq 0$. $F(z)$ is no longer an $\mathcal{O}$-transform and may not be a causal transform. Hence $F(z)$ is an entire function of order one not necessarily regular in a half-plane or even amenable to be made regular in a half plane by shift of origin.

We demonstrated (§ 2) that whatever the value of $\boldsymbol{a}$ ($|a| < \infty$) we have

$$h(0) = h(\pi) = 0,$$

but $h(\theta) \neq 0$ throughout a half plane unless $a = 0$. Therefore, for $a \neq 0$ we have

$$\lim_{r\to\infty} \frac{\ln|F(r\,\mathrm{e}^{\mathrm{i}\theta})|}{r} = Q, \tag{3.7}$$

where Q is a constant such that $Q \neq 0$ and $|Q| = |h(\theta)| \leqslant b < \infty$. This means $F(z)$ is behaving exponentially (either increasing ($a < 0$) or decreasing ($a > 0$), and so a function is required which goes to zero as $r \to \infty$. A corollary to Jordan's lemma shows that

$$\lim_{r\to\infty} \frac{\ln|F(r\,\mathrm{e}^{\mathrm{i}\theta})|}{r^{1+\epsilon}} = 0 \tag{3.8}$$

for $\epsilon > 0$. Therefore we define a new function containing $\ln F(z)$, which, with the assumptions made, will be a causal transform, by considering the convergent integral

$$\int_C \frac{\ln F(z)\,\mathrm{d}z}{z^{1+\epsilon}(z-x')}. \tag{3.9}$$

We choose the parameter ϵ as a positive integer; $\epsilon = 1$ is sufficient although $\phi(x)$ can still be obtained without difficulty for $\epsilon > 1$ (§3(*d*)). Thus:

$$\ln F(x') = -\frac{\mathrm{i}(x')^2}{\pi} P\int_{-\infty}^{\infty} \frac{\ln F(x)\,\mathrm{d}x}{x^2(x-x')} + x'\frac{\mathrm{d}}{\mathrm{d}x}(\ln F(x))\Big|_{x=0} + \ln F(0). \tag{3.10}$$

We let

$$\frac{\mathrm{d}}{\mathrm{d}x}(\ln F(x))\Big|_{x=0} = K$$

say, as the origin $x = 0$ can be chosen arbitrarily, one may always ensure that K is finite (Toll 1956).

From (3.10) we have

$$\phi(x') = -\frac{(x')^2}{\pi} P\int_{-\infty}^{\infty} \frac{\ln|F(x)|\,\mathrm{d}x}{x^2(x-x')} + x'\,\mathrm{Im}\,K + \phi(0). \tag{3.11}$$

From equation (3.11) we may find the phase from the modulus of $F(x)$, for any bandlimit (a, b) to within an arbitrary linear term. The numerical evaluation is complicated by the singularity at $x = 0$ (the singularity at $x = x'$ is avoided by use of the convolution theorem (von Fey 1956; Saxton 1974)). The problem at $x = 0$ is avoided by subtracting the value of $\ln|F(x)|$ at $x = 0$ (Toll 1956; Hilgevoord 1960) equation (3.9) then becomes

$$\int_C \frac{\ln F(z) - \ln F(0) - x'\dfrac{\mathrm{d}}{\mathrm{d}z}(\ln F(z))\Big|_{z=0}\,\mathrm{d}z}{z^2(z-x')} \tag{3.12}$$

which leads to the same result as (3.11).

(*d*) *Recursion relation between l.H.ts*

If $\ln F(z)$ behaves as a polynomial of degree n, for example when $F(z)$ is an entire function of finite type and of order n, then $\ln F(z)/z^{n+1}$ will satisfy the square integrability and regularity conditions of theorem 3.

Thus in general

$$\phi(x') = -\frac{(x')^{n+1}}{\pi} P\int_{-\infty}^{\infty} \frac{\ln|F(x)|}{x^{n+1}(x-x')}\,dx + K_1(x')^n + K_2(x')^{n-1} \ldots + \phi(0), \quad (3.13)$$

where K's are constants and n is an integer value of ϵ (see also Muskhelishvili 1953). We note that the Fourier transform of $z^{-n}\ln F(z)$ will vanish for negative arguments (from theorem 3), i.e. this function is a causal transform.

In practice, with a band limited $F(x)$, $F(z)$ is of finite type and of order 1. Hence $\ln F(z)$ behaves at most as an exponential in z in a half plane and the g.H.t. (equation (3.11)) is always sufficient to recover the phase though it may not be necessary. In fact, the m.H.t. may also be used, but it renders $F(z)$ into an $\mathcal{O}$-transform.

Provided that all the integrals which are implied below actually exist, we may write the general relation

$$P\int_{-\infty}^{\infty} \frac{g(x)\,dx}{x(x-x')} = \frac{\hat{g}(x')}{x'} - \frac{\hat{g}(0)}{x'}, \quad (3.14)$$

where $\hat{g}(x')$ is the Hilbert transform of $g(x)$. This equation shows that if the boundary conditions are chosen such that the correct phase is given by an r.H.t., the same phase is given by an m.H.t. (to within at most an arbitrary first degree polynomial). Furthermore, by replacing $g(x)$ in (3.14) by $x^{-n}g(x)$, a recursion relation may be derived establishing equation (3.11), the g.H.t., and showing that the phases agree to within the arbitrary polynomial.

In summary, we consider the contour integral

$$\int_C \frac{\ln F(z)\,dz}{z^n(z-x')} = 0,$$

where the contour C is the real axis and a semicircle, Γ, of infinite radius. For a function $F(x)$ defined by equation (2.1), the choice $n = 2$ (g.H.t.) is *always* sufficient to ensure that the integral along Γ does not contribute and that the integral along the real axis is convergent. The choice $n = 1$ (m.H.t) will *always* ensure that the integral along the real axis is convergent but leads to a zero contribution from the integral along Γ *only* if $a = 0$. The choice $n = 0$ (r.H.t.) does not result in a convergent integral along the axis unless $|F(x)|$ contains a d.c. level and even then the contribution from Γ vanishes only if $a = 0$.

4. The effect of zeros in the u.h.p.

Many functions will have zeros in more than a half plane and we examine the effect of these zeros on the phase determination and put forward methods for taking their contributions into account. Whenever $F(z) = 0$, $\ln F(z)$ has a branch point.

Evaluating equation (3.11) by a closed contour integration in the u.h.p. implies that the branch cuts were avoided, but their contribution to the integral must be included to give the actual phase $\phi(x)$. We must therefore add to (3.10) a term $\sum_j R_j(z)$ where the sum is over the j zeros which lie in the u.h.p. Considering the most general case and taking real and imaginary parts (of (3.9) for $\epsilon = 1$) we obtain

$$\ln|F(x')| = \frac{(x')^2}{\pi} P\int_{-\infty}^{\infty} \frac{\phi(x)\,\mathrm{d}x}{x^2(x-x')} + x' \operatorname{Re} K + \ln|F(0)| + 2(x')^2 \operatorname{Re} \sum_j R_j(x'), \quad (4.1)$$

and

$$\phi(x') = -\frac{(x')^2}{\pi} P\int_{-\infty}^{\infty} \frac{\ln|F(x)|\,\mathrm{d}x}{x^2(x-x')} + x' \operatorname{Im} K + \phi(0) + 2(x')^2 \operatorname{Im} \sum R_j(x'). \quad (4.2)$$

The term $R_j(z)$ is the functional contribution due to the jth zero at $z_j = x_j + iy_j$. From the measured modulus we evaluate the g.H.t. integral

$$\phi_{\mathrm{H}}(x') = \frac{(x')^2}{\pi} P\int_{-\infty}^{\infty} \frac{\ln|F(x)|\,\mathrm{d}x}{x^2(x-x')}; \quad (4.3)$$

(the m.H.t. might equally well have been used). We shall designate this phase the 'Hilbert phase' (sometimes it is known in the literature as the 'minimal phase' or 'canonical phase'). The function constructed from $|F(x)|$ and $\phi_{\mathrm{H}}(x)$ will be called the 'Hilbert function', $H(x)$, associated with $F(x)$.

From equation (4.2), the actual phase is given by

$$\phi_{\mathrm{H}}(x) = -\phi(x) + 2x^2 \operatorname{Im} \sum_j R_j(x) \quad (4.4)$$

neglecting linear terms and constants. To evaluate $\sum_j R_j(x)$ it is necessary to know the coordinates and the order of each zero which lies in the u.h.p. However, if the locations of the zeros are determined, rather than determining $\sum_j R_j(x)$, the function $F(z)$ can be directly determined (equation (4.6)).

We present first a method for eliminating the zeros from the u.h.p. Second, we discuss the means of calculating $F(z)$ by taking the zeros into account assuming them to be of order unity.

(*a*) *Removal of zeros from the u.h.p.*

If $\sum_j R_j(z) = 0$ for the (unmodified) function $F(z)$, the zeros lie on the real axis and/or in the l.h.p. and the actual phase $\phi(x)$ is given directly by the Hilbert phase $\phi_{\mathrm{H}}(x)$. This occurs for the complex degree of coherence of black body radiation (Wolf 1962) and the function $\operatorname{sinc} x$ (O'Neill & Walther 1963). Other examples are discussed by Titchmarsh (1939, 1948), Pólya (1918) and Cartwright (1955).

If $F(z)$ has zeros in the u.h.p. there is the possibility of removing them by modifying the function in a way which follows from Rouché's theorem (Holland 1973), which states: If two functions $F(z)$ and $R(z)$ are regular in a region C of the complex plane and if $|F(z)| < |R(z)|$ at every point of the boundary of C then this is a sufficient

but not necessary condition for $R(z)$ and $F(z)+R(z)$ to have the same number of zeros in C.

$R(z)$ is termed a reference function if it is chosen as having no zeros in the u.h.p. and satisfying Rouché's theorem in this half plane. Then the function $R(z)+F(z)$ has no zeros in the u.h.p. Hence the behaviour of $R(z)+F(z)$ is dominated at least from this point of view, by $R(z)$. The function $\mathscr{R}(z) = R(z)+F(z)$, zero free in the u.h.p., is therefore identical to its associated Hilbert function, i.e. its actual phase is the Hilbert phase. We shall call this particular type of Hilbert function a 'reference transform' ($\mathscr{R}$-transform). A function $F(z)$ with no zeros in the u.h.p. will be regarded as the simplest $\mathscr{R}$-transform, for which $R(z) \equiv 0$.

For Rouché's theorem to be verified, $R(z)$, must increase in the u.h.p. not slower than $F(z)$ when $a < 0$ and must not decrease faster than $F(z)$ for $a \geqslant 0$. Evidently, a suitable function for both cases will be $R(z) = A\,\mathrm{e}^{\mathrm{i}cz}$, where $c \not> a$ and $|A| > |F(x)|$ for all x.

In the particular case $a = 0$, the largest acceptable value for c becomes zero and this corresponds to $R(z) = A$, a constant. The addition of $R(z)$ to an analytic function moves the zeros but does not destroy them. This follows from an analytic function not having a local minimum (or maximum) in its modulus within its region of analyticity, except for minima at zeros. For example, in the important particular case of $F(z)$ being an $\mathscr{O}$-transform, if we choose $c = 0$, i.e. $R(z) =$ constant, then, since $F(z)$ tends to zero as $r \to \infty$ in the u.h.p. there must always be a contour in the u.h.p. along which any added constant, however small, will have a modulus greater than $|F(z)|$. This contour will mark the region where zeros may occur since by Rouché's theorem, there can be no zeros beyond this contour. As the constant is increased the contour moves across the real axis, i.e. all the zeros move into the l.h.p. This procedure makes it also possible to use the r.H.t. (§ 3 (*a*)) for obtaining the phase. The optimum constant, bearing in mind experimental error in the determination of $|F(x)|$ will produce a contour tangential to the real axis from the l.h.p. Practical applications of this condition are discussed in § 5 (*a*). It is never possible, strictly speaking, to add an $R(z)$ having the form $A\,\mathrm{e}^{\mathrm{i}cz}$ as this corresponds to a Dirac δ-function in the conjugate space, located to one side of the spectrum of $F(z)$; however, a reasonable approximation can be achieved.

(*b*) *Number and location of zeros*

In general, neither the position nor the order of the zeros is known. However, their location and order (which we assume to be unity) cannot be arbitrary since these determine the modulus on the real axis (Titchmarsh 1925, lemma 4.4) by the expression

$$|F(x)| = |F(0)| \prod_{j=1}^{\infty} \left| 1 - \frac{x}{r_j} \exp(-\mathrm{i}\theta_j) \right|, \tag{4.5}$$

where the jth zero is at $z_j = r_j\,\mathrm{e}^{\mathrm{i}\theta_j}$.

Zeros on the real axis do not cause any ambiguity, but if $F(z)$ has N zeros in the complex plane (not on the real axis), then replacing z_j by its complex conjugate does

not affect the modulus on the real axis. Thus 2^N possible solutions have this measured modulus, and so a reflection of an arbitrary number of zeros about the real axis is a source of ambiguity. A complete set of 2^N phases for a given modulus may be generated by 'zero flipping' which all give rise to functions with the same band limit (Walther 1963) because the density of the zeros is preserved. Since the Hilbert phase corresponds to a function with no zeros in the u.h.p., all the zeros from the u.h.p. of $F(z)$ are reflected into the l.h.p. to form the Hilbert function. The z_j values for the Hilbert phase function may be found by the analytic continuation of the Hilbert function (or the intensity, see, for example, Saxton 1975) in the l.h.p.

The Fourier transform of any bandlimited function has a denumerable infinity of zeros distributed throughout the complex plane (i.e. $z(x) \to \infty$, $x \to \infty$). As $x \to \infty$ the zeros tend asymptotically to lie along the real axis at the Nyquist frequency (Bond & Cahn 1958). Fortunately, in practice, the number of zeros that need be considered is finite, as will be discussed below.

The function $F(z)$, having zeros in the u.h.p., may be expressed as the product of two analytic functions (Titchmarsh 1939) as follows:

$$F(z) = H(z) \prod_{j=1}^{M} \left(\frac{z-z_j}{z-\bar{z}}\right), \tag{4.6}$$

where $$B_j(x) = \frac{H(x)\, y_j(y_j + \mathrm{i}(x - x_j))}{(x - x_j)^2 + y_j^2}$$

and $$A_{jl} = \frac{\mathrm{i}2y_l}{\bar{z}_j - \bar{z}_l}.$$

Taking the Fourier transform of (4.8) gives

$$f(t) = h(t) - 2 \sum_{j=1}^{N} k_j b_j(t) \prod_{\substack{l=1 \\ l \neq j}}^{N} (1 - k_l A_{jl}), \tag{4.9}$$

where $h(t)$ and $b_j(t)$ are Fourier transforms of $H(x)$ and $B_j(x)$. The similarities between equations (4.8) and (4.9) appear because the product term in each is a function only of the zero positions.

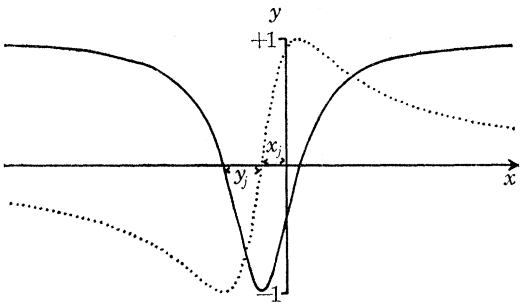

FIGURE 1. Typical form of a single term of the Blaschke product, on the real axis, for a zero at $z_j = x_j + \mathrm{i}y_j$. Full line real part, broken line imaginary part.

Equation (4.9) contains a sum of terms in k_j. The coefficient

$$c_j = -2k_j \prod_{\substack{l=1 \\ l \neq j}}^{N} (1 - k_l A_{jl})$$

of each term contains all the other k values and neither $b_j(t)$ nor c_j can ever be identically zero. If a system of simultaneous equations of the form (4.9) were solved, a zero jth term implies that the corresponding k_j is zero and a finite term necessarily means that the corresponding k_j is unity. In principle N independent measurements should allow the determination of the k_j values. (Since a non-zero c_j must have a modulus $\geqslant 1$, one would assume that any calculated c_j value significantly less than unity represents a zero value: see appendix.)

There are two obvious methods for solving the equations (4.8) and (4.9). The first makes use of a known, off-axis reference wave $r(t)$, which may not be suitable for the application of Rouché's theorem. If $r(t)$ is known, then for t values where $r(t)$ is spatially separate from $f(t)$ we may use this *a priori* information to solve equation (4.9) for the set of k_j values. If the addition of a reference is not possible, the second method (see § 5) relies on combining information from two experiments.

A typical term in the Blaschke product for the jth zero has the form shown in figure 1. Since y_j decreases asymptotically for increasing x_j (Titchmarsh 1925), zeros

with x_j values much larger than the experimental region of interest can be neglected since the corresponding term in the product approaches unity over this region. Similarly, for y_j large compared to the distance over which the solution is required, the corresponding term in the product tends to a constant and can be ignored. For a given minimum signal to noise ratio of recording $|F(x)|$, the area of the complex plane over which zero values need be considered is finite. These zeros lie in a strip parallel to the real axis of width dependent upon the variation of $f(t)$ within the support (Cartwright 1930, 1931).

5. Practical methods of phase determination

There are many existing methods which attempt to retrieve the phase from intensity measurements. Our aim in this section is to suggest their underlying unity as revealed by our approach. It is possible to demonstrate how each of the methods considered can be regarded as being conceptually equivalent to one of the ways of treating the zeros already outlined. The realization of this underlying unity may lead to the formulation of new experimental approaches. We distinguish between two categories of experimental approach, requiring either the effective removal of the zeros from the u.h.p., or the location of the zeros in the u.h.p. from equation (4.8) and/or (4.9). Within the second category there are two approaches:

(i) Using a known input (e.g. a known reference wave) with a set of measured data to solve equation (4.8) or (4.9).

(ii) Relating two sets of data *either* from measurements made in two spaces (Fraunhofer and image), *or* from two different, i.e. independent, sets of measurements made in one of these two spaces.

(*a*) *Removing the zeros from the u.h.p.*

The aim is to create an $\mathscr{R}$-transform making use of Rouché's theorem by experimentally utilizing a suitable reference function as defined in § 4(*a*). A possible function was given as $R(z) = \mathrm{A}\,\mathrm{e}^{\mathrm{i}cz}$, where $c \notin (a, b)$ and $|A| > |F(x)|$ for all x.

There are two ways in which such a reference function may be made available: one may either explicitly add the reference function or, in suitable cases, modify the unknown function $f(t)$, in such a way that it automatically contains its own reference function.

If $f(t)$ possesses a suitably well defined maximum within the support $a \leqslant t \leqslant b$, then a zero free half plane may be created by the use of a spatial filter, the strong maximum in $f(t)$ being used to provide the reference function. To illustrate this situation, let us consider the case of Gabor holography.

In Gabor holography the strong component lies at the centre of the t (diffraction) space and corresponds to the in-line reference beam. If we describe this reference beam by the function $R(x) = A\,\mathrm{e}^{\mathrm{i}cx}$, we have a situation where $b > c > a$ and in this spatial location the strong component cannot be used as a reference function since

we are unable to apply Rouché's theorem to either half plane. The same restriction applies if the strong component lies anywhere within the support of the function $f(t)$ since then also $b > c > a$. However, if a spatial filter is introduced which removes the spectrum to one side of this strong component at c, then we enforce the condition that c corresponds to one end of the support. If now $|A| > |F_1(x)|$ for all x, and $F_1(x)$ is the Fourier transform of the spatial frequency components not obstructed by the spatial filter and excluding the component at c, then $R(x)$ is a suitable reference function for $F_1(x)$ and so a zero free half plane, i.e. an $\mathscr{R}$-transform has been created. This approach was put forward, although for a different reason, by Bryngdahl & Lohmann (1968) and is known as single sideband (s.s.b.) holography. Another method removing part of the spectrum has been proposed by Misell, Burge & Greenaway (1974) and the Schlieren technique may also be regarded as a method falling in this category. However, part of the information about $F(x)$ has been lost and this may or may not be significant.

If such a strong component does not exist in $f(t)$, then a reference function must be added in the form of an off-axis reference beam. Using the function $A\,\mathrm{e}^{\mathrm{i}cx}$ for this reference beam, we require a sufficiently strong component, $|A| > |F(x)|$, which is positioned in such a manner that $c \not> a$. In t-space this corresponds to a strong component which occurs to one side of the spectrum of the function $F(x)$. This method was put forward, again for different reasons by Leith & Upatnieks (1964); it is the well known off-axis holography. The discussion of holography as a means of phase retrieval relying on the removal of the zeros from the half plane will be considered in terms of wavefront reconstruction.

Consider $\mathscr{R}(x) = F(x) + R(x)$, or equivalently $\mathscr{P}(t) = f(t) + r(t)$. Let the bandwidth of the function $f(t)$ be much greater than that of $r(t)$. The recorded intensity is $|\mathscr{R}(x)|^2$ and the Fourier transform of this intensity is $\tilde{\mathscr{P}}^2(t)$, the convolution square of $\mathscr{P}(t)$. Denoting the complex conjugate of $f(t)$ by $f^*(t)$, etc., we have:

$$\tilde{\mathscr{P}}^2(t) = f(t) \circledast f^*(-t) + r(t) \circledast r^*(-t) + f(t) \circledast r^*(-t) + r(t) \circledast f^*(-t), \qquad (5.1)$$

where $\circledast$ denotes convolution

We assume that $R(x)$ is known and investigate the separation of these terms as the reference beam is moved progressively off-axis. We shall compare the usual holographic decoding and the l.H.t. approach as a means for retrieving $F(x)$. Four important cases can be distinguished and these are shown in figure 2.

The case (i) shows all four terms from equation (5.1) overlapping and $f(t)$ cannot be determined from $\tilde{\mathscr{P}}^2(t)$. All the information concerning $F(x)$ is present but it is not directly accessible using hologram decoding because of the twin image problem. $F(x)$ is not an $\mathscr{R}$-transform and hence we cannot use directly an l.H.t. to determine it. $F(x)$ can be made an $\mathscr{R}$-transform however by the use of a spatial filter (s.s.b. holography). Alternatively a second experiment used in conjunction with this hologram might enable one to locate the zeros as indicated in § 4 (*b*) and § 5 (*b*).

Case (ii), off-axis holography, corresponds to an $\mathscr{R}$-transform for a sufficiently large $R(x)$ (in the sense defined in § 4 (*a*)); in this case $\mathscr{R}(x)$ is identical to its associated

Hilbert function and hence by applying an l.H.t., $F(x)$ may be determined. We note, however, that the holographic decoding cannot be applied (except as an approximation) to this situation since the cross-correlation terms overlap with the auto-correlation terms although no longer with each other. The same conclusions for case (ii) apply to case (iii). Its relevance as a distinct situation will become clear in § 5 (*b*).

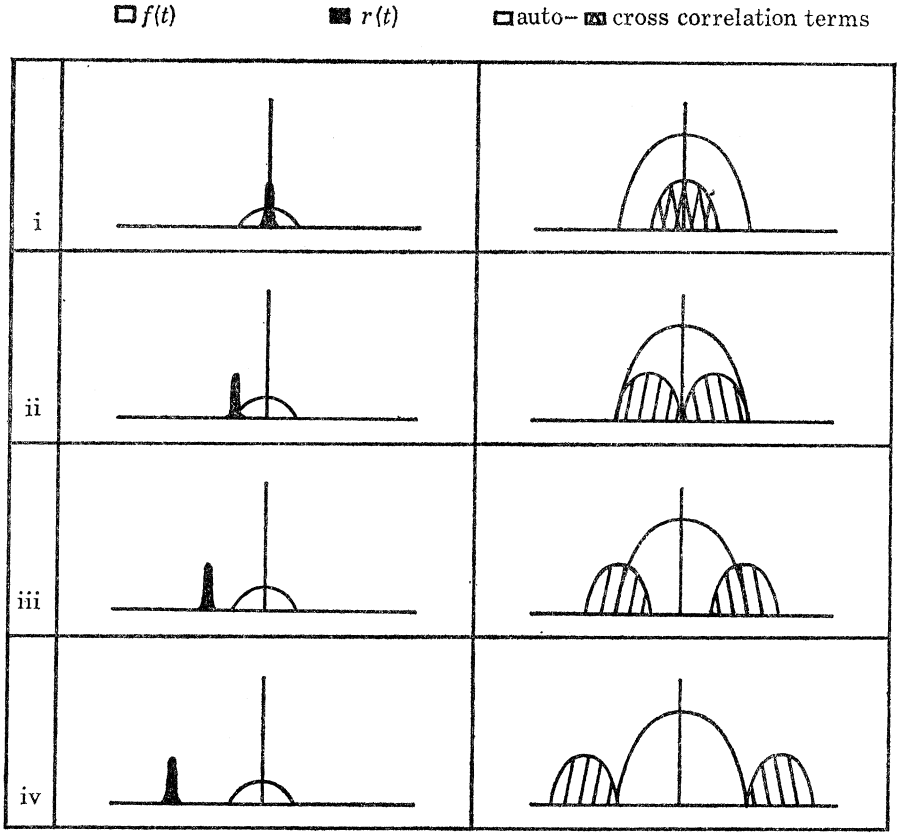

FIGURE 2. The Fourier transform of an $\mathscr{R}$-transform and its modulus, as the reference function is moved off-axis.

Finally, case (iv) represents 'true' off-axis holography which corresponds to distinct separation of the cross-correlation from the auto-correlation terms. In this situation both holographic decoding and the l.H.t. (again assuming $R(x)$ is sufficiently strong) could be used to retrieve the phase successfully.

As the reference beam moves further off-axis (case (ii) to case (iv)) a higher degree of temporal coherence is required due to the increasing path difference between the reference and scattered beams. This points towards a possible advantage in some experimental situations of using (ii) (with the l.H.t. decoding which it then requires), rather than (iv). In addition, it is noted that in case (ii) the frequency of the speckle

pattern is practically equal to that of the interference fringes while in case (iv) the frequency of the speckle pattern – equal to that in (ii) – is at most half that of the interference fringes.

(*b*) *Location of the zeros*

When an $\mathscr{R}$ transform cannot be created, or when there is no reference wave at all, it is necessary to locate the zeros of the function using the equations presented in § 4 (*b*).

We consider first the problem when we have a weak reference wave, not strong enough to produce an $\mathscr{R}$-transform, under the conditions depicted in (ii), (iii) and (iv) of figure 2. If we assume, as in § 5 (*a*), that $r(t)$ is known everywhere, then because at least part of $r(t)$ does not overlap with $f(t)$, we may solve equation (4.9) in order to find the k_j values. This may be understood as follows. Suppose that $r(t)$ is non-zero to the interval (a, b). It is known that zero flipping will not change the overall bandwidth (a, b). Thus the information that the spectrum is zero outside this range, cannot be used to solve equation (4.9). However, for t values in the range $a \leqslant t < d$ we may replace the left hand side of (4.9) with a known function. Thus, for these t values, we may solve (4.9) for the c_j's (for example, by Gauss elimination). The experimental validity of true off-axis holography as a method gives confidence that a unique solution of equation (4.9) exists in at least case (iv). Equation (4.9) shows that when the number of zeros, N, corresponding to the available experimental range is small, the relevant set of k_j's may even be determined by inspection of the coefficients $b_j(t)$ and A_{jl}. Unless there exists a high degree of symmetry in $\mathscr{P}(t)$, (in which case Schiske-type ambiguities (1974) are possible), it seems reasonable to conjecture that the solution will also be unique for cases (ii) and (iii): note, however, that case (iv) can never be symmetric.

Some information about any symmetry in the position of the zeros can also be obtained by using other *a priori* knowledge as has been suggested by Bates (1969) and Roman & Marathay (1963).

Secondly when the reference wave is strong but not sufficiently off-axis to satisfy Rouché's theorem, the alternative to the possibility of spatial filtering discussed in § 5 (*a*) is to solve the equations (4.8) and/or (4.9) to determine the k_j's. Two independent sets of data are required in order to achieve a solution. One procedure proposed by Frank (1973) relies on two measurements: one with the central spot in the scattering pattern (Gabor holography, i.e. a bright field image), and one without the central spot (a dark-ground image). In relation to the analysis presented here there are two sets of equations to be solved of the form (4.8), in terms of $F(x) + A$, (A assumed known), and $F(x)$ respectively.

Lohmann (1974) has given a method of phase determination based on the suppression in two separate experiments, of the information in two small and different ranges of t-space. The success of this method can again be interpreted in terms of the provision of sufficient information to characterize the k_j's, the only difference being

the replacement of the known constant A by a known function. There are other experimental techniques which rely on the procedure of modifying the spectrum of $F(x)$ (see, for example, Misell 1973; Frank 1972).

When two planes are physically accessible it becomes possible to measure $|F(x)|$ and $|f(t)|$ and so solve for the k_j's from a combination of equations (4.8) and (4.9). Unfortunately, taking the modulus of (4.9), as is required in this method leads to computational complications. This idea is contained implicitly in the method of phase recovery put forward by Gerchberg & Saxton (1972). These authors do not determine the zero positions explicitly but use a Fourier iterative scheme between the two data sets. This method is an alternative to using equations (4.8) and (4.9) but lacks a procedure (Schiske 1974) for judging the uniqueness of the resulting phase distribution.

6. Summary and discussion

In an physically conceivable situation objects and instruments are of finite dimensions, so that the application of the theory of entire functions will always be justified for phase retrieval. We have presented this theory only for the one dimensional case. A summary of all the theoretical cases discussed in this paper with their consequences, is provided in the form of a flow chart (figure 3), which illustrates the important points.

There will always be some difficulties in practical application of the theory. One limitation is that only finite intervals are available in each space. This implies that the Hilbert transform calculated will be an approximation to the theoretical expression defined on an infinite interval.

Phase retrieval procedures applicable to realizable physical objects may all be regarded as providing on $\mathscr{R}$-transform or sufficient data to determine the zero locations in the u.h.p. by solving equation (4.8) or (4.9). For a large number of zeros we expect that solving these equations will be tedious and in the presence of noise the reliability of the solution obtained will be questionable. The removal of the zeros from the u.h.p. by the creation of an $\mathscr{R}$-transform obviates the need to solve equations (4.8) and (4.9), since in this case the actual phase is the Hilbert phase. We emphasize that, in spite of the fact that a true plane wave can never be obtained in a physical situation, the $\mathscr{R}$-transform approach is far preferable to any method of zero location. Any Gaussian-like approximation to a plane wave may be expected to present large areas adjacent to the real axis which are zero free (Nussenzveig 1967). When using such approximations as reference functions, one must ensure that Rouché's theorem is satisfied in a sufficiently large area adjacent to the region of experimental interest. This is the situation used in practical holography when the reference wave is invariably taken to be plane and strong with respect to the unknown wave.

It would be particularly interesting to investigate in detail the classes of functions which have a zero free half plane, since it may then become possible to identify a

general class of objects for which the Hilbert phase is the actual phase. The equality (e.g. Bond & Cahn 1958) between the number of zeros and the number of degrees of freedom, may reflect a deeper connection between the two concepts. We feel that this is worthy of further study.

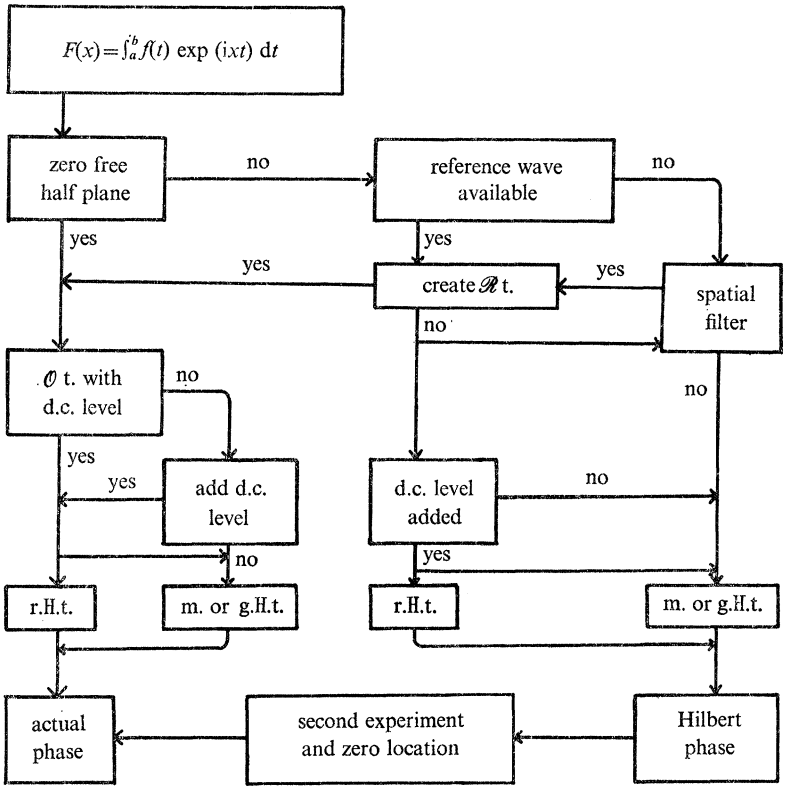

FIGURE 3. A flow chart illustrating the steps required to solve the phase problem.

With respect to causality (see footnote on page 194 and Wolter (1961)) one would expect in spatial problems that such a requirement would be unreasonable as it may be regarded as obliterating half of the information present. Nevertheless, the Hilbert transform which relates the real and imaginary parts of a function does require that the causality condition, i.e. the requirement of single-sidedness, is fulfilled. In spatial problems, however, the observable magnitude is always the modulus of the measured function and the requirement of single sidedness need not apply to the functions of interest but only to the functions suitably modified for use with the Hilbert relations. The causality requirement does not restrict in any way the value of the upper limit b (equation (2.1)) which, from the constraints of Titchmarsh's theorem (theorem 3) need not be finite. However, if b were infinite the order of the function need not be unity and the associated indicator function may not be defined. This

would imply that the derivation and subsequent application of the m.H.t. or g.H.t. could not be justified in any obvious way, if at all.

Two of us (M. A. F. and A. H. G.) acknowledge Science Research Council research studentships. S.R.C. support for the work is also acknowledged.

Appendix

Consider one term in the product in equation (4.6) for the particular case of a zero at z_1 therefore

$$P_l(z) = \frac{z - z_l}{z - \bar{z}_l}$$

and

$$P_l(x) = 1 - \frac{2y_l(y_l + \mathrm{i}(x - x_l))}{(x - x_l)^2 + y_l^2}$$

$$= 1 - 2y_l\,\delta(x - x_l) \circledast \left(\frac{y_l + \mathrm{i}x}{x^2 + y_l}\right), \tag{A 1}$$

where $\delta(x)$ is the Dirac δ-function. The Fourier transform of (A 1) is

$$p_l(t) = \delta(t) - 2y_l\,\mathrm{e}^{-\mathrm{i}x_l t}\,\mathrm{e}^{-y_l t}\,H(t), \tag{A 2}$$

where

$$H(t) = 1 \quad (t > 0)$$
$$= 0 \quad (t < 0) \quad \text{(Champeney 1973)}.$$

A similar expression to (A 2) can be written for a zero at z_m. Convolving $p_l(t)$ and $p_m(t)$ gives

$$p_l(t) \circledast p_m(t) = \delta(t) - 2y_l\,\mathrm{e}^{-\mathrm{i}\bar{z}_l t} - 2y_m\,\mathrm{e}^{-\mathrm{i}\bar{z}_m t} + 4y_l y_m(I), \tag{A 3}$$

where

$$I = \mathrm{e}^{-\mathrm{i}\bar{z}_l t}\,H_l \circledast \mathrm{e}^{-\mathrm{i}\bar{z}_m t}\,H_m$$
$$= \mathrm{e}^{-\mathrm{i}\bar{z}_m t}\left[\frac{\mathrm{e}^{-\mathrm{i}(\bar{z}_l - \bar{z}_m)t} - 1}{-\mathrm{i}(\bar{z}_l - \bar{z}_m)}\right]$$
$$= \frac{\mathrm{i}(\mathrm{e}^{-\mathrm{i}\bar{z}_l t} - \mathrm{e}^{-\mathrm{i}\bar{z}_m t})}{\bar{z}_l - \bar{z}_m}.$$

Hence

$$p_l(t) \circledast p_m(t) = \delta(t) - 2y_l\left(1 - \frac{\mathrm{i}2y_m}{\bar{z}_l - \bar{z}_m}\right)\mathrm{e}^{-\mathrm{i}\bar{z}_l t} - 2y_m\left(1 - \frac{\mathrm{i}2y_l}{\bar{z}_m - \bar{z}_l}\right)\mathrm{e}^{-\mathrm{i}\bar{z}_m t}.$$

Similarly for a further convolution by $p_n(t)$ one obtains

$$p_l(t) \circledast p_m(t) \circledast p_n(t) = \delta(t) - 2y_l\left(1 - \frac{\mathrm{i}2y_n}{\bar{z}_l - \bar{z}_n}\right)\left(1 - \frac{\mathrm{i}2y_m}{\bar{z}_l - \bar{z}_m}\right)\mathrm{e}^{-\mathrm{i}\bar{z}_l t}$$
$$- 2y_m\left(1 - \frac{\mathrm{i}2y_n}{\bar{z}_m - \bar{z}_n}\right)\left(1 - \frac{\mathrm{i}2y_l}{\bar{z}_m - \bar{z}_l}\right)\mathrm{e}^{-\mathrm{i}\bar{z}_m t}$$
$$- 2y_n\left(1 - \frac{\mathrm{i}2y_l}{\bar{z}_n - \bar{z}_l}\right)\left(1 - \frac{i2y_m}{\bar{z}_n - \bar{z}_m}\right)\mathrm{e}^{-\mathrm{i}\bar{z}_n t}.$$

In general, therefore,

$$p_1(t) \circledast \ldots p_N(t) = \delta(t) - 2 \sum_{j=1}^{N} y_j \mathrm{e}^{-\mathrm{i}\bar{z}_j t} \prod_{\substack{l=1 \\ l \neq j}}^{N} \left(1 - \frac{\mathrm{i}2y_l}{\bar{z}_j - \bar{z}_l}\right). \tag{A 4}$$

In order to include all the zeros, the parameters k_j are introduced such that $k_j = 1$ if the zero is in the u.h.p. and $k_j = 0$ if the zero appears in the l.h.p. Zeros on the real axis are detectable and so are not included here. Thus, Fourier transforming equation (A 4) we have

$$\prod_{j=1}^{N} P_j(x) = 1 - 2 \sum_{j=1}^{N} k_j y_j \left(\frac{y_j + \mathrm{i}(x - x_j)}{(x - x_j)^2 + y_j^2}\right) \prod_{\substack{l=1 \\ l \neq j}}^{N} \left(1 - \frac{\mathrm{i}2y_l k_l}{\bar{z}_j - \bar{z}_l}\right),$$

which is equation (4.8).

Using the definitions following equations (4.8) and (4.9) (§ 4(*b*)), these two equations can be written

$$F(x) = H(x) + \sum_{j=1}^{N} B_j(x)\, c_j, \tag{A 5}$$

$$f(t) = h(t) + \sum_{j=1}^{N} b_j(t)\, c_j, \tag{A 6}$$

where

$$c_j = -2k_j \prod_{\substack{l=1 \\ l \neq j}}^{N} (1 - k_l A_{jl}). \tag{A 7}$$

Now

$$|(1 - k_l A_{jl})|^2 = 1 + \frac{4k_l y_l y_j}{|z_j - z_l|^2} \geqslant 1. \tag{A 8}$$

The c_j values are finite and we may rewrite equations (4.8) and (4.9) in the form (A 5) and (A 6). No c_j value may be zero unless the corresponding k_j is identically zero. Further, a non-zero c_j must have modulus $\geqslant 1$. Thus, the problems of phase retrieval is reduced to a yes/no problem with regard to the location of zeros in the u.h.p.: a finite c_j value implies yes, a zero value implies no. If a c_j is finite then $|c_j| \geqslant 1$. Approximate solutions to (A 5) and (A 6) may be sought and any calculated c_j which is significantly less than unity may be taken to be a zero value.

References

Bates, R. H. T. 1969 *Mon. Not. Roy. astr. Soc.* **142**, 413.

Boas, R. P. 1954 *Entire functions.* New York: Academic Press.

Bond, F. E. & Cahn, C. R. 1958 *I.R.E. Trans. Inf. Theory*, p. 110.

Bryngdahl, O. & Lohmann, A. 1968 *J. Opt. Soc. Am.* **58**, 620.

Burge, R. E., Fiddy, M. A., Greenaway, A. H. & Ross, G. 1974 *J. Phys. D* **7**, L65.

Cartwright, M. L. 1930 *Q. Jl Math.* (Oxford Series) (1) **1**, 38.

Cartwright, M. L. 1931 *Q. Jl Math.* (Oxford Series) (1) **2**, 113.

Cartwright, M. L. 1955 *Integral functions.* Cambridge Tracts in Mathematics and Mathematical Physics. No. 44. Cambridge University Press.

Champeney, D. C. 1973 *Fourier transforms and their physical applications.* London: Academic Press.

Fey, M. von 1956 *Nachrichtentechnik, Berlin* **8**, 337.
Frank, J. 1972 *Biophys. J.* **12**, 484.
Frank, J. 1973 *Optik* **38**, 582.
Gerchberg, R. W. & Saxton, W. O. 1972 *Optik* **35**, 237.
Goedecke, G. H. 1975 *J. Opt. Soc. Am.* **65**, 146.
Hilgevoord, J. 1960 *Dispersion relations and causal description*. Amsterdam: North Holland.
Hoenders, B. J. 1975 *J. Math. Phys.* **16**, 1719.
Holland, A. S. B. 1973 *Introduction to the theory of entire functions*. London: Academic Press.
King, G. I. 1975 *Acta Cryst.* A **31**, 130.
Leith, E. N. & Upatnieks, J. 1964 *J. Opt. Soc. Am.* **54**, 1295.
Lohmann, A. W. 1974 *Optik* **41**, 1.
Misell, D. L. 1973 *J. Phys.* D **6**, 2200.
Misell, D. L., Burge, R. E. & Greenaway, A. H. 1974 *J. Phys.* D **7**, L27.
Muskhelishvili, N. I. 1953 *Singular integral equations*. Groningen: P. Noordhoff.
Nussenzveig, H. M. 1967 *J. Math. Phys.* **8**, 561.
Nussenzveig, H. M. 1972 *Causality and dispersion relations*. London: Academic Press.
O'Neill, E. L. & Walther, A. 1963 *Opt. Acta* **10**, 33.
Page, C. H. 1955 *Physical mathematics*. New York: Van Nostrand.
Pěrina, J. 1971 *Coherence of light*. London: Van Nostrand.
Pólya, G. 1918 *Math. Z.* **2**, 353.
Roman, P. & Marathay, A. S. 1963 *Il nuovo cim.* **30**, 1452.
Schiske, P. 1974 *Optik* **40**, 261.
Saxton, W. O. 1974 *J. Phys. D* **7**, L63.
Saxton, W. O. 1975 Ph.D. Thesis, University of Cambridge.
Toll, J. S. 1956 *Phys. Rev.* **104**, 1760.
Titchmarsh, E. C. 1925 *Proc. Lond. Math. Soc.* (2) **25**, 283.
Titchmarsh, E. C. 1939 *The Theory of Functions*, 2nd ed. Oxford University Press.
Titchmarsh, E. C. 1948 *The theory of the Fourier integral*, 2nd ed. Oxford University Press.
Van Kampen, N. G. 1953 *Phys. Rev.* **89**, 1072.
Walther, A. 1963 *Opt. Acta* **10**, 41.
Wolf, E. 1962 *Proc. Phys. Soc.* **80**, 1269.
Wolter, H. 1961 *Progress in optics* **1**, 157.