Huygens's wave propagation principle corrected

David A. B. Miller

AT&T Bell Laboratories, Crawfords Corner Road, Holmdel, New Jersey 07733

Received June 10, 1991

Huygens's principle that each point on a wave front represents a source of spherical waves is conceptually useful but is incomplete; the backward parts of the wavelets have to be neglected *ad hoc*, otherwise backward waves are generated. The problem is solved mathematically by Kirchhoff's rigorous integration of the wave equation, but the intuitive appeal of Huygens's simple principle is lost. I show that, by using spatiotemporal dipoles instead of spherical point sources, one can recover a simple principle of scalar wave propagation that is correct whenever the concept of a wave front is meaningful.

Huygens's principle[1] that every point on a wave front can be considered as a new source of spherical wavelets is a powerful conceptual tool in understanding wave propagation. When combined by Fresnel[2] with the principle of interference, this concept explained key phenomena of diffraction. Both Huygens and Fresnel had to neglect backward parts of the wavelets arbitrarily, however. Otherwise, wave propagation in free space cannot be described properly, because backward waves are generated. The problem for scalar waves was solved rigorously by Helmholtz for the monochromatic case[3] and by Kirchhoff more generally.[4] The result can be expressed through Kirchhoff's integral theorem, to which the Huygens–Fresnel approach can be shown to be an approximation. The standard interpretation of Kirchhoff's surface integral terms involves two types of sources of varying strengths, so the simplicity of Huygens's approach is lost. Kirchhoff then approximated his own rigorous result to obtain his useful diffraction formula, in which the wavelets become progressively weaker, for angles θ to the normal to the wave front, by an inclination factor $1 + \cos\theta$ (see, e.g., Refs. 5 and 6). There is, however, no simple physical source that can give rise to these wavelets. Hence Huygens's original idea that wave propagation can be described in terms of simple effective sources on a wave front appears not to work. Here, however, I demonstrate that Huygens's concept does work, and is rigorously correct, provided that we use spatiotemporal dipoles rather than Huygens's original point sources.[7] This principle of scalar wave propagation[8] is valid whenever the concept of a wave front is meaningful.

Consider a scalar wave equation

$$\nabla^2\phi - \frac{1}{c^2}\frac{\partial^2\phi}{\partial t^2} = -g(\mathbf{r}, t), \tag{1}$$

where ϕ is the scalar wave amplitude, c is the velocity of propagation, and $g(\mathbf{r}, t)$ is a source term. Suppose that there are sources only within a volume V bounded by a surface S. Then we can use the techniques of Kirchhoff's integration of the wave equation (see, e.g., Ref. 9) to prove that, for $\mathbf{r}_1$ outside V,

$$\phi(\mathbf{r}_1, t) = \frac{1}{4\pi}\int_S \left\{-\frac{1}{r}\left[\frac{\partial\phi}{\partial n}\right] + [\phi]\frac{\partial}{\partial n}\left(\frac{1}{r}\right) - \frac{1}{cr}\times\frac{\partial r}{\partial n}\left[\frac{\partial\phi}{\partial t}\right]\right\} da. \tag{2}$$

Here $r = |\mathbf{r}_1 - \mathbf{r}_0|$ is the distance from point $\mathbf{r}_1$ to the point $\mathbf{r}_0$ of interest on the surface S in the integral, and n is the distance parallel to the outward normal to the surface S. The square brackets denote a quantity evaluated at the retarded time $t - r/c$.

The meaning of the surface integral is well known. Instead of having actual sources of waves inside the volume V, we could have exactly the same wave for all points outside V (and zero inside V) if we had an appropriate set of real sources on the surface S. The value and type of these sources are given by the terms in the integrand in Eq. (2). The first term represents point sources of spherical waves. The second and third terms together represent spatial dipoles[10] (or doublets[5]) oriented perpendicular to the surface.

If, however, we restrict S to being a wave front, we can reinterpret these surface sources. If we choose three orthogonal Cartesian directions, $\hat{\mathbf{n}}$ (the normal to the surface), $\hat{\mathbf{q}}_2$, and $\hat{\mathbf{q}}_3$, at any point on the surface, we can define a wave front[11] as a closed surface on which

$$\frac{\partial^2\phi}{\partial n^2} \gg \frac{\partial^2\phi}{\partial {q_2}^2}, \frac{\partial^2\phi}{\partial {q_3}^2}. \tag{3}$$

This condition [relation (3)] allows us to approximate the wave equation near the surface by a one-dimensional wave equation with propagation in the direction $\hat{\mathbf{n}}$. Hence the wave propagation is locally perpendicular to the wave front as required. The general solution of such a wave equation corresponding to an outward propagating wave is $f(n - ct)$, where f is an arbitrary function. Hence, near the surface,

$$\frac{\partial\phi}{\partial n} = -\frac{1}{c}\frac{\partial\phi}{\partial t}. \tag{4}$$

0146-9592/91/181370-03$5.00/0

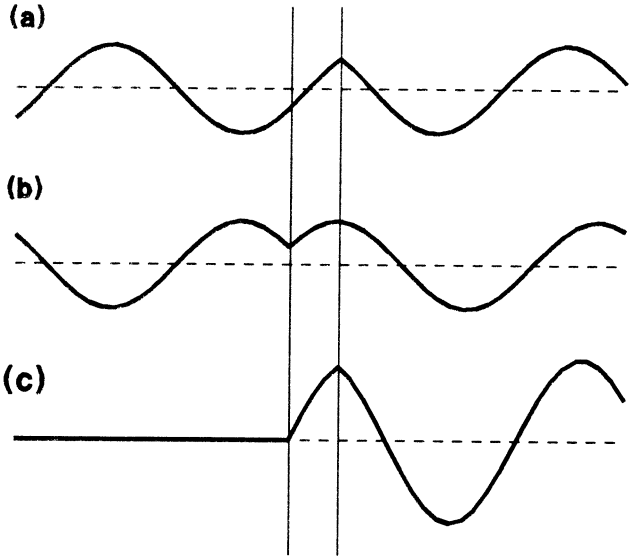


Fig. 1. Wave from a spatiotemporal dipole for plane wave propagation for the approximate case of finite separation of the dipole. The position of the two sources is indicated by the vertical lines. The sources are separated by a distance d. (a) Waves propagating out from the right source. (b) Waves propagating out from the left source, which is delayed by a time $\tau = d/c$ compared with the right source. (c) Net wave from the two sources. Note that there is no resulting wave in the left direction, and the waves do not cancel on the right-hand side.

For a monochromatic wave with time factor $\exp(i\omega t)$, and with $k = \omega/c$, the condition [relation (3)] that we put on the surface S becomes

$$\frac{\partial^2\phi}{\partial q_2^2}, \frac{\partial^2\phi}{\partial q_3^2} \ll k^2\phi, \tag{5}$$

which is equivalent to saying that the change in $\partial\phi/\partial q$ over a wavelength λ is much less than ϕ/λ, in agreement with our intuitive picture of a wave front. Incidentally, the model presented here is actually exact for uniform spherical or plane wave fronts, both being ideal wave fronts.

For our monochromatic wave, we now have, from Eq. (4) $\partial\phi/\partial n = -ik\phi$, so that the integrand in Eq. (2) becomes

$$\frac{1}{4\pi}\left\{-\frac{1}{r}\left[\frac{\partial\phi}{\partial n}\right] + [\phi]\frac{\partial}{\partial n}\left(\frac{1}{r}\right) - \frac{1}{cr}\frac{\partial r}{\partial n}\left[\frac{\partial\phi}{\partial t}\right]\right\} = \frac{[\phi]}{4\pi r}\left\{ik(1 + \cos\theta) + \frac{\cos\theta}{r}\right\}, \tag{6}$$

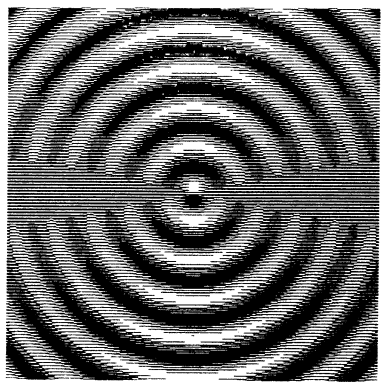
where we have used the fact that $\partial r/\partial n = -\cos\theta$, since θ is the angle between $\hat{\mathbf{n}}$ and $\mathbf{r}_1 - \mathbf{r}_0$ and r decreases as we move the source point $\mathbf{r}_0$ along $\hat{\mathbf{n}}$.

To understand the meaning of the integrand as given in Eq. (6), consider the following spatiotemporal dipole source. This source consists of (i) a point source of strength $+a$, located at point $d/2$ on the $\hat{\mathbf{n}}$ axis, and (ii) another point source, of strength $-a$, located at point $-d/2$ on the $\hat{\mathbf{n}}$ axis, delayed with respect to the first source by a time $\tau = d/c$, i.e., delayed by the time taken for a wave to propagate from the first to the second source. [Here the sign convention is that a positive source corresponds to positive g in Eq. (1).] We will call this a spatiotemporal dipole of strength ad. The wave from a single (monochromatic) point source of strength $+a\exp(i\omega t)$ is $(a/4\pi r)\exp[i(\omega t - kr)]$, where r is the distance from the source. Hence simple algebra shows that the resulting wave from this spatiotemporal dipole, in the limit as d goes to zero (but with ad remaining finite), is

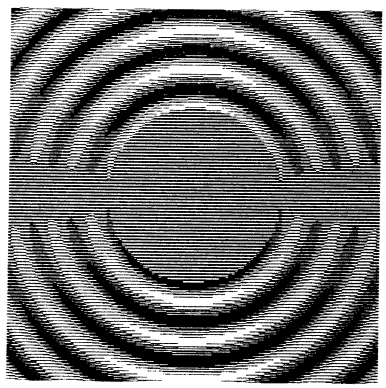
$$\phi_{TD} = \frac{ad\, e^{i(\omega t - kr)}}{4\pi r}\left\{ik(1 + \cos\theta) + \frac{\cos\theta}{r}\right\}. \tag{7}$$

However, this is exactly the same as the integrand in Eq. (2), as expressed in Eq. (6), provided only that the strength of the spatiotemporal dipoles is set to $ad\exp(i\omega t) = \phi$ per unit area. {Note that $\phi\exp(-ikr) = [\phi]$.}

Hence we come to the new wave propagation principle: For a closed wave front S, the monochromatic wave propagation outside S is equivalent to that from a set of spatiotemporal dipoles, oriented perpendicular to the surface S, and of strength ϕ



(a)



(b)

Fig. 2. (a) Exact wave from a (spatial) dipole radiator. (b) Wave calculated using the proposed wave propagation principle, using spatiotemporal dipoles on a spherical surface of radius 2.25 wavelengths. The amplitude of the spatiotemporal dipoles per unit area is chosen equal to the wave amplitude on the spherical wave front of radius 2.25 wavelengths in (a). The dipoles are oriented perpendicular to the surface. For visual clarity, the wave amplitude is multiplied by the distance from the center to remove the underlying $1/r$ dependence. A seven-level gray scale is used. In the region outside the chosen wave front, both waves are seen to be similar.

per unit area on the surface S, where ϕ is the wave amplitude on S.

Essentially, then, we have simply substituted spatiotemporal dipoles, oriented perpendicular to the wave front, for Huygens's original point sources. At first sight, the spatiotemporal dipole concept is unusual. It is, however, easy to understand for plane waves (Fig. 1). The spatiotemporal dipoles can now be approximated by two separate sets (or sheets) of sources, separated by some distance d, in which the left-hand set has the opposite sign and is delayed by a time $\tau = d/c$ compared with the right-hand source. In the limit as $d \to 0$, this becomes exact. Each source [Figs. 1(a) and 1(b)] generates both left- and right-propagating waves as shown. Note, however, that the delay is such that the left-propagating wave from the left source exactly cancels the left-propagating wave from the right source, which gives no net wave in the left direction [Fig. 1(c)]. On the right-hand side, by contrast, the two waves do not cancel, so that the net effect is only a right-propagating wave. Spatiotemporal dipoles also have several other interesting properties.[12]

For direct numerical illustration, in Fig. 2 I have chosen the wave from an oscillating (spatial) dipole, which is shown exactly in Fig. 2(a). In Fig. 2(b), I show the calculated wave that results for a sphere of spatiotemporal dipole sources oriented perpendicular to the sphere surface; the strengths per unit area are given by the wave amplitude on the wave front of radius 2.25 wavelengths in Fig. 2(a). Outside the chosen wave front, the calculated waves are almost identical, even though I have chosen an extreme example in which the amplitude on the wave front is not slowly varying (changing from a positive maximum to a negative maximum in a circumferential distance of $2.25\pi \sim 7$ wavelengths). The actual error is $\lesssim 7\%$ of the peak amplitude on a given wave front for this relatively extreme case. Note that the spatiotemporal dipole sources produce essentially no backward wave, as required, with essentially no wave amplitude inside the chosen wave front. Again, the wave inside the chosen wave front is not exactly zero because this is an approximation, but it is too small to be visible in Fig. 2(b). Incidentally, Kirchhoff's approximate diffraction formula, which corresponds to dropping the near-field term $\cos\theta/r^2$ of Eq. (6), does not correctly predict the wave near the chosen wave front. The current principle is, however, valid in the near field, as can be seen in Fig. 2, and can be used in the sense of Huygens's original wave propagation idea to calculate one wave front from the effective sources on the previous wave front.

The point of the above calculation is to show explicitly that the proposed principle works, not to suggest that this is an efficient way to calculate wave propagation. The aim of this principle is to understand wave propagation conceptually; it remains to be seen whether it helps directly in actual calculations.

In conclusion, I have recovered a simple picture of scalar wave propagation, much like Huygens's original notion, but in which one uses spatiotemporal dipoles oriented perpendicular to the wave front instead of Huygens's simple point sources. With this one correction, this principle now encompasses Fresnel's and Kirchhoff's mathematical models for all cases where the concept of a wave front is meaningful.

I am pleased to acknowledge many helpful comments and corrections from the reviewers of this Letter.

References

1. C. Huygens, *Traité de la Lumière* (Leyden, 1690) [English translation by S. P. Thompson, *Treatise on Light* (Macmillan, London, 1912)].
2. A. Fresnel, Ann. Chem. Phys. **1,** 239 (1816); Mem. Acad. **5,** 339 (1826).
3. H. von Helmholtz, J. Math. **57,** 7 (1859).
4. G. Kirchhoff, Berl. Sitzungsber. **641** (1882); Ann. Phys. **18,** 663 (1883); Vorlesungen über Math. Phys. **2** (1891).
5. M. Born and E. Wolf, *Principles of Optics*, 6th ed. (Pergamon, Oxford, 1980).
6. B. B. Baker and E. T. Copson, *Mathematical Theory of Huygens' Principle*, 2nd ed. (Oxford U. Press, London, 1950).
7. Part of this research was presented by D. A. B. Miller, in *Digest of Optical Society of America Annual Meeting* (Optical Society of America, Washington, D.C., 1990), paper PD16.
8. Note that such scalar solutions, while valid, e.g., for acoustic waves, are not complete solutions for electromagnetic waves because Maxwell's equations impose additional constraints (see Ref. 6).
9. See, for example, J. A. Stratton, *Electromagnetic Theory* (McGraw-Hill, New York, 1941).
10. The spatial dipole, consisting of equal and opposite infinitesimally separated point sources, should not be confused with a radiating electric dipole, which is a source of vector electromagnetic waves.
11. Some waves (e.g., the net wave from two discrete point sources) cannot be adequately described using wave fronts; despite this, the wave front is, however, undoubtedly useful conceptually in understanding waves.
12. Other features of the spatiotemporal dipoles include the following: (i) Although only considered here for the monochromatic case, the spatiotemporal dipoles also work for the general time-dependent case. (ii) There are two different kinds of spatiotemporal dipoles, the second kind having the opposite relative delay; this second kind corresponds to waves propagating inward rather than outward. (iii) Kirchhoff's surface integral can be rewritten exactly in terms of the two kinds of spatiotemporal dipoles instead of point and doublet sources; thus, instead of having surface sources to set the overall wave amplitude (doublets) and its normal derivative (point sources), we can use the two kinds of spatiotemporal dipoles to set the outward and inward wave amplitudes, respectively.