Forum for Electromagnetic Research Methods and Application Technologies (FERMAT)

The Demise of the Hundred Year Old Mythology of the Famous Sommerfeld Sign Error Along with a Realization of the Zenneck wave and its relationship with Surface, Lateral and Leaky waves

Tapan K. Sarkar
Dept. of Electrical Engineering and Computer Science,
Syracuse University, Syracuse, NY, 13244-1240 USA.

Magdalena Salazar-Palma
Departmento de Teoría de la Senal y Comunicaciones,
Universidad Carlos III de Madrid, 28911 Leganes, Madrid, Spain

Abstract—**The famous mythology of the sign error in Sommerfeld formulation is debunked. It is amazing that this misleading information about the sign error has continued for almost a hundred years with researchers stating that there is a sign error in the Sommerfeld formulation but nobody actually was able to show exactly where the sign error occurred. In addition, Sommerfeld never accepted to the sign error myth and also this mythical error was not found even when a group of scientists from AT&T bell labs went through Sommerfelds's paper with a fine tooth comb. The only person who went through Sommerfeld's formulation very methodically and categorically stated that there was no error in Sommerfeld's work was first Wise and Rice whose works were quoted by Stratton in his book. In recent years, Robert Collin also demonstrated that the pole contribution in the Sommerfeld integrals is cancelled by the branch cut contribution and so no surface wave term shows up in the final solution. Then Collin conjectures that the reason for the discrepancy can be use of different approximations for the expansion of the complementary error function. Following Collin's suggestion it is shown that indeed based on the approximation of the complementary error function one can get a surface wave term or different terms like a lateral wave which is a collection of leaky waves and alternatively the ground wave depending on what type of approximation is made to the Sommerfeld integrals. It is also illustrated that a surface wave does not generally arise in a two media problem with a single boundary layer unless the dielectric permittivity changes sign across the boundary and the imaginary part is much lower than the magnitude of the negative real part. In addition, the history of evolution of the Zenneck wave, the surface wave, the leaky wave and the lateral waves are described. A realization of the Zenneck wave is illustrated for cellular wireless communication using the classical Sommerfeld theory where it may also manifest itself as a lateral wave. Surface waves are generated in a three media problem where a trapped wave is generated in one of the layers. Also clearer demarcation is made between Zenneck waves which are not true surface wave and a surface wave. Finally leaky waves and the lateral waves are described. All these waves are defined based on where the pole of the Green's function occur on which Riemann sheet and therefore they are function of the chosen branch cut as is illustrated here and hence they do not have a clear crisp definition as to how they are categorized.**

1. INTRODUCTION AND DISCUSSION OF THE EARLY HISTORY

Electromagnetics started with Faraday and quantified by Maxwell [1]. As presented by Wait [2] and Collin [3], in the 19th century, after the initial work of Hertz [4], and Heaviside [5], and exemplified by Lodge [6] and followed through by Tesla [7], Popov [8], Bose [9], and others, researchers had irrefutable evidence of the correctness of Maxwell's equations and the ability to generate and radiate electromagnetic waves. However, it was Marconi [10] who translated all these principles into reality by transmitting electromagnetic waves over long distances and proved to the scientific community that besides the line-of-sight communication other modes of communication were possible! The conjecture at that time was that perhaps the transmission over such large distances took place through the *surface wave* introduced earlier by Lord Rayleigh [11].

According to Schelkunoff [11] it was Lord Rayleigh who discovered that in a semi-infinite elastic medium a source of finite dimensions excites two kinds of waves: 1) *space waves* which spread in all directions and 2) *surface waves* which spread only along the boundary. If the medium is non-dissipative, it follows from the principle of conservation of energy that at large distances from the source, the energy density in a space wave varies inversely as the square of the distance from the source, while in a surface wave it varies inversely as the distance. Surface waves seemed to be attached to the boundary of the solid and tended to follow it even if it was curved.

In the time of Marconi's famous experiments and prior to the discovery of the Kennelly-Heaviside reflect-ing layer, there was much speculation about possible existence of similar kinds of electromagnetic waves. It was already known that electric waves had a tendency to cling to parallel wires ("Lecher wires," as they were called) and thus could be guided around corners. Did the surface of the earth have a similar tendency to cap-ture some of the energy from an antenna and guide it into the shadow, thus explaining Marconi's success? That was the question. In the half-century that followed the question, the answer had been at first "Yes," then "No," and a controversy began.

This created a great interest in the study of surface waves. The guiding of a plane electromagnetic wave along the flat interface separating air and the imperfectly conducting ground seems to have been first investigated by Cohn [12] and shortly thereafter by Uller [13]. Zenneck [14,15] recognized the bearing of these researchers on the propagation of radio waves and showed that the field equations admit a solution that can be interpreted as a surface wave guided by a plane interface separating any two media.

As further stated by Schelkunoff, according to Zenneck [15] "there are waves which emanate from a transmitter placed in a homogenous insulating material and energy is radiated in straight lines, radially from the transmitter. Consequently, the energy varies as $1/\rho^2$ (ρ= distance from the source) and the amplitudes of the electric and magnetic field strengths vary as $1/\rho$, which are termed as space waves. A different kind of wave is obtained for an antenna located at the earth's surface. The wave emanated into the air by an antenna at the earth's surface may be conceived as consisting of two parts, one of which is of the nature of a space wave and the other of a surface wave. In the former, the energy $\propto \frac{1}{\rho^2}$, the amplitude therefore varies as $\propto \frac{1}{\rho}$; in the latter, the energy $\propto \frac{1}{\rho}$, and therefore the amplitude $\propto \frac{1}{\sqrt{\rho}}$. The fact that in the latter there is a decrease in energy as the distance increases in contrast to a wave following a wire – and in addition to and entirely aside from such absorption as occurs – is explained by the fact that the energy is spreading itself out over in ever-increasing circles as the wave propagates. This much is relevant to the classical distinction between space and surface waves. Therefore, initially Zenneck and Sommerfeld [15,16] accepted Rayleigh's definition of a surface wave as far as the most significant physical properties are concerned. However, later they have made an unfortunate slip in their analysis which subsequently confused the issue [11]. The controversy thus arose from the following statement of Zenneck [15]: "While at short distances from the transmitter, the waves are almost entirely of the nature of space waves, as the distance increases the surface component becomes more and more predominant, as its amplitude decreases more slowly than that of the space component. That is the nature of the wave constantly approaches that of a surface wave. When the distance becomes very great, the surface wave may again give way to the space wave, as the former is more rapidly absorbed. It is questionable, however whether this effect is of practical importance."

The above quoted conclusion of Zenneck's is based on the original formulas obtained by Sommerfeld. However, it was found later on both from theory and experiment that the surface wave term is missing from the complete final solution initially proposed by Sommerfeld. Therefore at this point we describe the various formulations of Sommerfeld and what were the relevant issues.

2. ORIGINAL SOMMERFELD FORMULATION FOR A VERTICAL ELECTRIC DIPOLE OVER AN IMPERFECT GROUND PLANE [16-19]

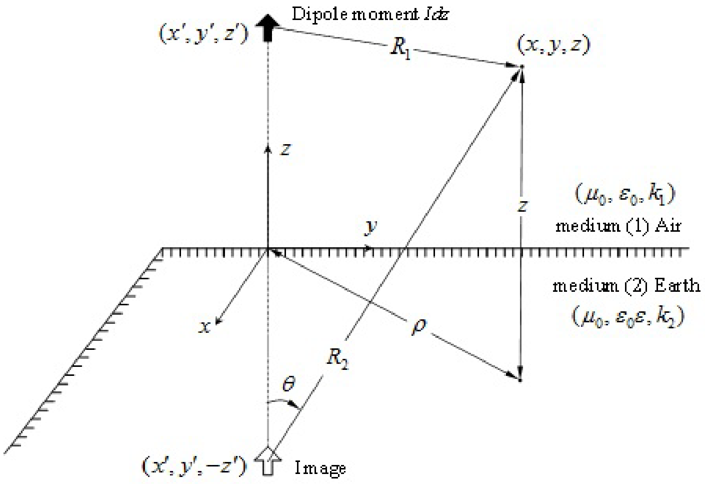


Figure 1 A vertical dipole over a horizontal imperfect ground plane.

Consider an elementary electric dipole of moment *Idz* oriented along the *z*-direction and located at (x', y', z'). The dipole is situated over an imperfect ground plane characterized by a complex relative dielectric constant *ε* as seen in Figure 1. The complex relative dielectric constant is given by $\varepsilon = \varepsilon_r - \frac{j\sigma}{\omega\varepsilon_0}$ where ε_r represent the relative permittivity of the medium, ε_0 is the permittivity of vacuum, σ is the conductivity of the medium, ω stands for the angular frequency, and j is the imaginary unit, i.e., $j = \sqrt{-1}$. It is possible to formulate a solution to the problem of radiation from the dipole operating in the presence of the imperfect ground in terms of a single Hertzian vector Π_z of the electric type. A time variation of $\exp(j\omega t)$ is assumed throughout the analysis, where, *t* is the time variable. The Hertzian vector $\hat{u}_z\Pi_z$ in this case satisfies the wave equation

$$\left(\nabla^2 + k_1^2\right)\Pi_{1z} = \frac{-I\,dz}{j\omega\varepsilon_0}\delta\left(x-x'\right)\delta\left(y-y'\right)\delta\left(z-z'\right) \tag{1}$$

$$\left(\nabla^2 + k_2^2\right)\Pi_{2z} = 0 \tag{2}$$

where

$$k_1^2 = \omega^2\mu_0\varepsilon_0 \tag{3}$$

$$k_2^2 = \omega^2\mu_0\varepsilon_0\varepsilon \tag{4}$$

and δ represents the delta function in space. The primed and unprimed coordinates are for the source and field points respectively. The subscript 1 denote the upper half space which is air and the subscript 2 denotes the lower half space which is the imperfectly conducting earth characterized by a complex relative dielectric constant ε. The electric and the magnetic field vectors are derived using the Hertzian vector from

$$\vec{E}_i = \vec{\nabla}\left(\vec{\nabla}\bullet\overrightarrow{\Pi_i}\right) + k_i^{\,2}\overrightarrow{\Pi_i} \tag{5}$$

and

$$\overline{H}_i = j\,\omega\,\varepsilon_0\,\varepsilon_i\left(\vec{\nabla}\times\overrightarrow{\Pi_i}\right) \tag{6}$$

respectively, with i = 1, 2. In medium 1, $\varepsilon_1 = 1$ and for medium 2, $\varepsilon_2 = \varepsilon$. So that the propagation constants in medium 1 and 2, called k_1 and k_2, and are related by $\frac{k_2}{k_1} = \sqrt{\varepsilon}$. At the interface $z = 0$, the tangential electric and magnetic field components must be continuous, conditions which in terms of the Hertzian vector components can be written as

$$\frac{\partial\Pi_{1z}}{\partial y} = \varepsilon\frac{\partial\Pi_{2z}}{\partial y} \tag{7a}$$

$$\frac{\partial\Pi_{1z}}{\partial x} = \varepsilon\frac{\partial\Pi_{2z}}{\partial x} \tag{7b}$$

$$\frac{\partial}{\partial y}\left(\frac{\partial\Pi_{1z}}{\partial z}\right) = \frac{\partial}{\partial y}\left(\frac{\partial\Pi_{2z}}{\partial z}\right) \tag{7c}$$

$$\frac{\partial}{\partial x}\left(\frac{\partial\Pi_{1z}}{\partial z}\right) = \frac{\partial}{\partial x}\left(\frac{\partial\Pi_{2z}}{\partial z}\right) \tag{7d}$$

Since all the boundary conditions must hold at $z = 0$ for all x and y, the x and y dependence of the fields on either side of the interface must be the same. Therefore

$$\Pi_{1z} = \varepsilon\,\Pi_{2z} \tag{8a}$$

$$\frac{\partial\Pi_{1z}}{\partial z} = \frac{\partial\Pi_{2z}}{\partial z} \tag{8b}$$

The complete solutions for the Hertz vectors satisfying the wave equations (1) and (2) and the boundary conditions (8) have been derived by many researchers over the last century. A partial list [16-33] that will be important to our discussions is provided starting with Sommerfeld [16]. The solutions are

$$\Pi_{1z} = P\left[\frac{\exp\left(-jk_1R_1\right)}{R_1} + \int_0^\infty \frac{J_0\left(\xi\,\rho\right)}{\sqrt{\xi^2-k_1^2}}\,\frac{\varepsilon\sqrt{\xi^2-k_1^{\,2}}-\sqrt{\xi^2-k_2^{\,2}}}{\varepsilon\sqrt{\xi^2-k_1^{\,2}}+\sqrt{\xi^2-k_2^{\,2}}}\exp\left(-\sqrt{\xi^2-k_1^{\,2}}\left(z+z'\right)\right)\xi\,d\xi\right] \tag{9}$$

and

$$\Pi_{2z} = 2P\int_0^{\infty} \frac{J_0(\xi\rho)\exp\left(\sqrt{\xi^2-k_2{}^2}\,z-\sqrt{\xi^2-k_1{}^2}\,z'\right)}{\varepsilon\sqrt{\xi^2-k_1{}^2}+\sqrt{\xi^2-k_2{}^2}}\,\xi\,d\xi \tag{10}$$

for $Real\left[\sqrt{\xi^2-k_{1,2}^2}\right] > 0$. $J_0(x)$ represents the zero-th order Bessel function of the first kind of argument x. Here

$$P = \frac{I\,dz}{j\omega 4\pi\,\varepsilon_0} \tag{11}$$

$$\rho = \sqrt{(x-x')^2+(y-y')^2} \tag{12}$$

$$R_1 = \sqrt{\rho^2+(z-z')^2} \tag{13}$$

and ξ is the variable of integration. For Π_{1z}, the first term inside the brackets of (9) can be interpreted as the particular solution or the direct line-of-sight (LOS) contribution from the dipole source and the second term can be interpreted as the complementary solution or a reflection term (reflection from the imperfect ground plane). We will call this potential responsible for the fields of the *ground wave*, as per IEEE Standard Definitions of Terms for Radio Wave Propagation [35]. Observe that the second term of this potential for the ground wave in (9) is the strongest near the surface of the earth and exponentially decays as we go away from the interface.

Similarly, the solution for Π_{2z} can be interpreted as a partial transmission of the wave from medium 1 into medium 2. With these thoughts in mind Π_{1z}, or equivalently the potential responsible for the ground wave, can be split up into two terms

$$\Pi_{1z} = \Pi_{1z}^{direct} + \Pi_{1z}^{reflected} = P\left(g_0+g_s\right) \tag{14}$$

where

$$\Pi_{1z}^{direct} = P\exp\left(-jk_1R_1\right)/R_1 = P\,g_0 \tag{15}$$

$$\Pi_{1z}^{reflected} = P\int_0^{\infty}\frac{\varepsilon\sqrt{\xi^2-k_1^2}-\sqrt{\xi^2-k_2{}^2}}{\varepsilon\sqrt{\xi^2-k_1{}^2}+\sqrt{\xi^2-k_2{}^2}}\,\frac{J_0(\xi\rho)\exp\left[-\sqrt{\xi^2-k_1{}^2}\,(z+z')\right]}{\sqrt{\xi^2-k_1{}^2}}\,\xi\,d\xi = Pg_s \tag{16}$$

The path of integration for the semi-infinite integral is labeled C_2 and is depicted in Figure 2 along with the singularities of the multivalued function, two branch points at $\pm k_1$ and $\pm k_2$ locations, and a pole p arising from the ratio of two functions in (16).

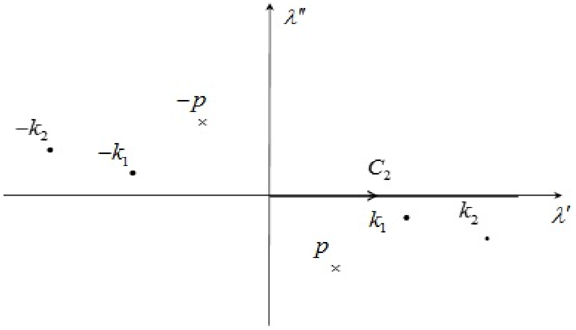


Figure 2: The contour of integration along the real axis from 0 to ∞ in the complex λ-plane.

A physical explanation to the two components of the Hertz potential Π_{1z} can now be given. The first one Π_{1z}^{direct} can be explained as a spherical wave originating from the source dipole. This term is easy to deal with. The difficult problem lies in the evaluation of $\Pi_{1z}^{reflected}$. Therefore, $\Pi_{1z}^{reflected}$ can be interpreted as a superposition of plane waves resulting from the reflection of the various plane waves into which the original spherical wave has been expanded and multiplied by a different value for the reflection coefficient for each ray. This arises from the identity

$$\frac{\exp\left(-j\,k_1 R_2\right)}{R_2}=\int_0^\infty \frac{J_0\left(\xi\rho\right)\exp\left[-\sqrt{\xi^2-k_1^{\,2}}\left(z+z'\right)\right]}{\sqrt{\xi^2-k_1^{\,2}}}\xi\,d\xi \tag{17}$$

for $\operatorname{Re}\left[\sqrt{\xi^2-k_1^2}\right]>0$ and

$$R_2=\sqrt{\rho^2+\left(z+z'\right)^2} \tag{18}$$

The reflection coefficient $R(\xi)$ is then defined as

$$R\left(\xi\right)=\frac{\varepsilon\sqrt{\xi^2-k_1^{\,2}}-\sqrt{\xi^2-k_2^{\,2}}}{\varepsilon\sqrt{\xi^2-k_1^{\,2}}+\sqrt{\xi^2-k_2^{\,2}}} \tag{19}$$

where the semi-infinite integral over ξ in $\Pi_{1z}^{reflected}$ takes into account all the possible plane waves. As $\varepsilon\to\infty$, i.e., a perfect conductor for the earth, then g_s of (16) reduces to (17) and represents a simple spherical wave originating at the image point. This physical picture will later be applied in the derivation of the reflection coefficient method. The reflection coefficient takes into account the effects of the ground plane in all the wave decomposition of the spherical wave and sums it up as a ray originating from the image of the source dipole but multiplied by a specular reflection coefficient $R(\theta)$, where θ is interpreted as the angle of the incident wave to the ground.

It is now important to point out that there are two forms of $\Pi_{1z}^{reflected}$ that may be used interchangeably as the two expressions are mathematically identical in nature (but have different asymptotic properties as we shall see) and are defined as

$$\begin{aligned}\Pi_{1z}^{reflected}&=P\left[\frac{\exp\left(-jk_1R_2\right)}{R_2}-2\int_0^\infty\frac{\sqrt{\xi^2-k_2^{\,2}}}{\sqrt{\xi^2-k_1^{\,2}}}\frac{J_0\left(\xi\rho\right)\exp\left[-\sqrt{\xi^2-k_1^{\,2}}\left(z+z'\right)\right]}{\varepsilon\sqrt{\xi^2-k_1^{\,2}}+\sqrt{\xi^2-k_2^{\,2}}}\xi\,d\xi\right]\\&=P\left[g_1-g_{sV}\right]\end{aligned} \tag{20}$$

where g_1 represents the spherical wave originating from the image of the source, and g_{sV} represents the correction factor to accurately characterize the effects of the ground. Equivalently, one can rewrite the same expression as

$$\begin{aligned}\Pi_{1z}^{reflected}&=P\left[-\frac{\exp\left(-jk_1R_2\right)}{R_2}+2\,\varepsilon\int_0^\infty\frac{J_0\left(\xi\rho\right)\exp\left[-\sqrt{\xi^2-k_1^{\,2}}\left(z+z'\right)\right]}{\varepsilon\sqrt{\xi^2-k_1^{\,2}}+\sqrt{\xi^2-k_2^{\,2}}}\xi\,d\xi\right]\\&=P\left[-g_1+G_{sV}\right]\end{aligned} \tag{21}$$

Now the image from the source has a negative sign along with the correction factor. This expansion is useful when both the transmitter and the receiver are located close to the ground and the reflection coefficient is −1 for grazing angle of incidence where $\theta\approx\pi/2$. Then the direct term g_0 cancels the image term g_1 leaving only the correction factor G_{sv}. Now for grazing incidence the fields are obtained using G_{sv}.

In order to evaluate the semi-infinite integrals, first the Bessel function of the first kind and zeroth order is transformed to a Hankel function of the first and second kinds and zeroth order through the use of the following identity by using

$$J_0(x) = \frac{1}{2}\left[H_0^{(1)}(x) + H_0^{(2)}(x)\right] \tag{22a}$$

and, also utilizing

$$H_0^{(1)}\left(xe^{j\pi}\right) = -H_0^{(2)}(x) \tag{22b}$$

where $H_0^{(1)}$ and $H_0^{(2)}$ are the Hankel functions of zeroth order and of first and second kinds, respectively. The semi-infinite integrals in (20) and (21) can be transformed to $-\infty$ to ∞. In the evaluation of (20) and (21), the crux of the problem lies in the characterization of the various branch points and singularities associated with (19) as illustrated in Figure 3 along with the locations of the branch cuts and the poles. One must also observe that ε is a complex quantity and its square root need to be taken with the proper sign.

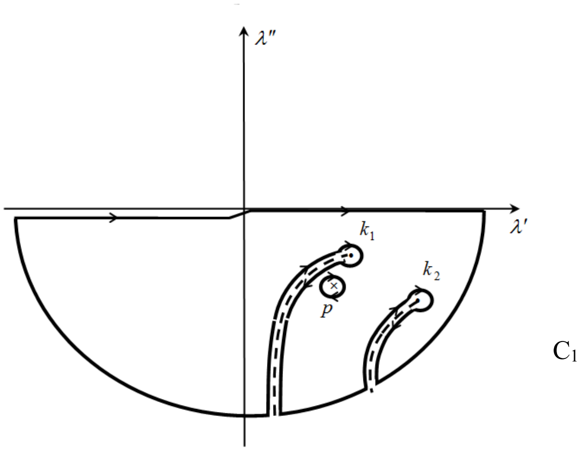


Figure 3: Actual location of the pole in the lower complex λ-plane and the contour (the outside enclosed line) C_1 enclosing them.

The first point to observe is that the second term of the Hertz potential denoted by a complex integral and particularly $R(\xi)$ has four branch points located at $\pm\, k_1$ and $\pm\, k_2$. Associated with these branch points are four branch cuts and this give rise to four Riemann sheets. On the four separate Riemann sheets, the following conditions are satisfied:

Sheet 1: $\mathrm{Re}\left(\sqrt{\xi^2 - k_1^2}\right) > 0$ and $\mathrm{Re}\left(\sqrt{\xi^2 - k_2^2}\right) > 0$ (23a)

Sheet 2: $\mathrm{Re}\left(\sqrt{\xi^2 - k_1^2}\right) < 0$ and $\mathrm{Re}\left(\sqrt{\xi^2 - k_2^2}\right) > 0$ (23b)

Sheet 3: $\mathrm{Re}\left(\sqrt{\xi^2 - k_1^2}\right) > 0$ and $\mathrm{Re}\left(\sqrt{\xi^2 - k_2^2}\right) < 0$ (23c)

Sheet 4: $\mathrm{Re}\left(\sqrt{\xi^2 - k_1^2}\right) < 0$ and $\mathrm{Re}\left(\sqrt{\xi^2 - k_2^2}\right) < 0$ (23d)

Sheet 1 is the proper Riemann sheet. Now, the function described in (19) has two zeros corresponding to the zeros of the numerator, usually called zeros associated with the Brewster's angle, and two poles corresponding to the zeros of the denominator, and usually called the surface wave poles. So, the zero of the reflection coefficient in (19) illustrates the Brewster's phenomenon [35] (i.e., the wave goes into the second medium for a particular angle of incidence without reflecting any energy) and an infinite value for the reflection coefficient illustrates the presence of a surface wave (i.e., a wave propagating close to the interface). The zeros and poles occur exactly at the same location $\xi_S = \pm\frac{k_1 k_2}{\sqrt{k_1^2 + k_2^2}}$. However on some Riemann sheets they appear as poles and on other Riemann sheets as zeros. So, the four poles and zeros are distributed on the four Riemann sheets. The most confusing stuff is that on the proper Riemann sheet whether it will be a pole or zero depends on the value of the dielectric constant ε. For example examining the denominator of $R(\xi)$ one can observe that for real values of ε on the proper Riemann sheet there is a zero at ξ_s. Whereas for a complex value of the dielectric constant ε, that zero may become a pole on the proper Riemann sheet. Therefore unless one specifies the values for the dielectric constant and chooses the proper Riemann sheet, it is difficult to know whether the pole will occur or not in the expression of (19). We first deal with the integral in (20) and then discuss about the form in (21).

The convergence of the integrals in (20) and (21) is assured even in the presence of the Hankel function, when $\operatorname{Im}\left(\xi\right)\right) \le 0$ as $\rho \to \infty$. Since $H_0^{(2)}\left(\xi\,\rho\right)$ can be integrated through the origin $\left(\xi = 0\right)$, the path C_2 of Figure 2 can now be modified to the path C_1 of Figure 3, following the real axis from $-\infty$ to $+\infty$. Now (20) can be rewritten as

$$g_{sV} = -j\int_{C_1} \frac{\sqrt{k_2^{\,2}-\xi^2}}{\sqrt{k_1^{\,2}-\xi^2}} \frac{\mathrm{H}_0^{(2)}\left(\xi\,\rho\right)\exp\left[-j\sqrt{k_1^{\,2}-\xi^2}\left(z+z'\right)\right]}{\varepsilon\sqrt{k_1^{\,2}-\xi^2}+\sqrt{k_2^{\,2}-\xi^2}}\,\xi\,d\xi \tag{24}$$

where the contour of integration C_1 is shown in the Figure 3 along with the location of the branch points at k_1 and k_2 and their associated branch cuts, together with the pole of (19).

The presence of the free term ξ in (24) will nullify the singularity of the Hankel function at $\xi = 0$. The integral from $-\infty$ to 0 goes slightly below the negative real axis as the Hankel function has a branch cut along that line. g_{sV} is now a spectrum of plane waves travelling away from the ground plane with the vertical component of the propagation constant as $\sqrt{k_1^{\,2}-\xi^2}$. The integral in (24) also contains double valued functions $\sqrt{k_{1,\,2}^{\;2}-\xi^2}$. The proper sheet of the double valued functions are those on which the radiation condition [p. 257, 26]

$$\left(\frac{\partial g_{sV}}{\partial z} - j\sqrt{k^2-\xi^2}\,g_{sV}\right)e^{-j\sqrt{k_1^{\,2}-\xi^2}z}\Big|_{z=-\infty}^{\infty} = 0 \tag{25}$$

is satisfied and can occur only if $\operatorname{Im}[\sqrt{k_1^{\,2}-\xi^2}] < 0$, i.e., $g_{sV} \to 0$ as $\rho \to \infty$ for a fixed z, and $g_{sV} \to 0$ as $z \to \infty$ for a fixed ρ. For large values of ρ, the asymptotic representation for the Hankel function is used

$$\mathrm{H}_0^{(2)}\left(\xi\,\rho\right)\xrightarrow[|\xi\,\rho|\to\infty]{\text{Limit}}\sqrt{\frac{2}{\pi\,\xi\,\rho}}\exp\left[-j\xi\,\rho + j\,{}^{\pi}\!/_{4}\right] \tag{26}$$

It is now important to point out that there are two forms of $\Pi_{1z}^{reflected}$ that may be used interchangeably as the two expressions in (20) and (21) are mathematically identical in nature (but have different asymptotic properties as we shall see).

Next, the following substitutions are made:

$$\xi = k_1 \sin\beta \tag{27}$$

$$\rho = R_2 \sin\theta \tag{28}$$

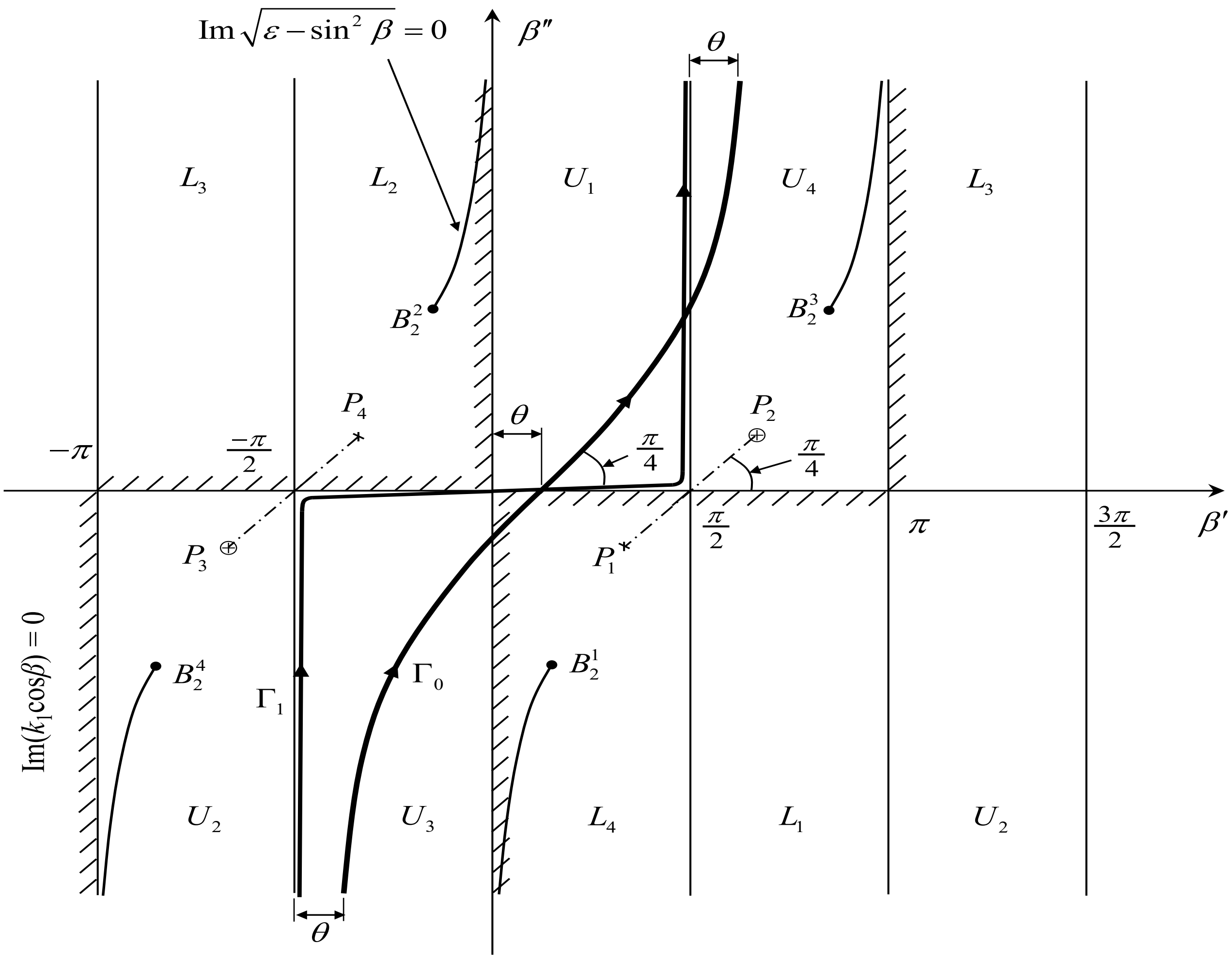
$$z + z' = R_2 \cos\theta \tag{29}$$

The interpretation of the angle θ is shown in Figure 1. Hence, the application of (27)-(29) to (24) yields

$$g_{sV} \approx \int_{\Gamma_1} \left[\frac{2k_1 \sin\beta}{\pi R_2 \sin\theta}\right]^{1/2} \frac{\sqrt{\varepsilon - \sin^2\beta}}{\varepsilon\cos\beta + \sqrt{\varepsilon - \sin^2\beta}} \times \exp\left[j\left\{-\pi/4 - k_1 R_2 \cos\left(\beta - \theta\right)\right\}\right] d\beta \tag{30}$$

where Γ_1 is a path in the complex β plane as shown in Figure 4. An approximate methodology utilizing the ordinary saddle point integration methodology is used to evaluate these infinite integrals. This is outlined in Appendix B. There is one obvious weakness in the arguments presented to derive (30), namely, that there are points on the path for which the argument of the Hankel function $H_0^{(2)}(\xi\rho)$ used in (24) is not large and may even be zero, so that the asymptotic expansion for large arguments cannot be used. However, as argued by Brekhovskikh [27], the arguments will be rigorous if the large argument approximation is used only after the path of integration has been changed to the path of steepest descent Γ_0 of Figure 4. The result will then be the same.

The mapping of (30), $\xi = k_1 \sin\beta$, transforms the quadrants of the ξ plane in parallel strips of width $\pi/2$ radians, and the path of integration from $\xi = -\infty$ to ∞ is transformed to the path Γ_1. The script '*U*' in Figure 4, denotes the strips of the β-plane to upper Riemann sheet. The other strips are denoted by 'L' (lower Riemann sheet). The path Γ_1 then totally lies on *U*. The location of the branch points at $\xi = \pm k_1$ in the ξ-plane are now transformed into $\sin\beta_{B_1} = \pm 1$ in the β-plane and are situated at $\pm\pi/2$, $\pm 3\pi/2$ and so on. The branch cuts along $\operatorname{Im}\left(\sqrt{k_1^2 - \xi^2}\right) = 0$ are now transformed to $\operatorname{Im}\left(k_1 \cos\beta\right) = 0$, and begin at the branch points labelled B_2^i located at $\beta = \pm\pi/2$. Since $\operatorname{Im}\left(k_1 \cos\beta\right) = \operatorname{Im}\left[k_1 \sin\left(\pi/2 \pm \beta\right)\right] = 0$, these branch cuts will run parallel to the path Γ_1, $\left[\operatorname{Im}\left(k_1 \sin\beta\right) = \lambda'' = 0\right]$ but shifted by $\pm\pi/2$ from the origin along the real axis. So the transformation $\xi = k_1 \sin\beta$ has transformed the upper and lower sheets associated with the branch points $\xi = \pm k_1$ into one sheet where certain strips on the sheet belong to the previous upper (U) and lower (L) Riemann sheet on the ξ-plane.



$\mathrm{Im}(k_1\cos\beta) < 0$ inside the shaded region

Figure 4: The complex β plane showing possible branch points, branch cuts, poles and the path of steepest descent for an imperfect ground plane with the material parameter $\varepsilon=\varepsilon'(1-j)^2$ and $|\varepsilon| >> 1$. The shaded region is $U_1 - U_4$.

The remaining branch points $\xi=\pm k_2$ are transformed to $\sin\beta_{B_2}=\pm\sqrt{\varepsilon}$ which has solutions

$$\beta_{B_2}=j\ln\left[\pm j\sqrt{\varepsilon}\pm\sqrt{1-\varepsilon}\right]. \tag{31}$$

The branch cuts $\mathrm{Im}\left[\sqrt{k_2^2-\lambda^2}\right]=0$ are transformed into $\mathrm{Im}\left[\sqrt{\varepsilon-\sin^2\beta}\right]=0$. In the β-plane there are then two Riemann sheets connected along the branch cuts $\mathrm{Im}\left[\sqrt{\varepsilon-\sin^2\beta}\right]=0$ of the branch points β_{B_2}. Finally the poles in the ξ-plane are now given by $\varepsilon\cos\beta_P+\sqrt{\varepsilon-\sin^2\beta_P}=0$. Hence,

$$\sin\beta_P=\pm\sqrt{\frac{\varepsilon}{1+\varepsilon}} \qquad \text{with} \qquad \cos\beta_P=-\sqrt{\frac{1}{1+\varepsilon}} \tag{32}$$

since $\mathrm{Im}\left(k_1\cos\beta\right)<0$. The possible locations for the poles can be approximated by

$\pm\sin\beta_P = \sqrt{\frac{\varepsilon}{1+\varepsilon}} \approx 1-\frac{1}{2\varepsilon} \approx \cos\left(\frac{\pi}{2}\mp\beta_P\right)$ which results in

$$\beta_P \approx \pm\left(\frac{\pi}{2}\mp\frac{1}{\sqrt{\varepsilon}}\right) \tag{33}$$

For the parameters of a highly conducting ground $\varepsilon_r = \varepsilon'(1-j)^2$ the locations of the poles and the branch cuts are pictorially depicted in Figure 4 (not to scale). Out of the possible locations of the branch points and poles, B_2^3, B_2^4, P_2 and P_3 are situated on the upper Riemann sheet of the branch points β_{B_2} on which $\operatorname{Im}\left(k_1\cos\beta\right) < 0$. It is also important to note that none of the poles (P_2, P_3) are situated between the original path of the integration Γ_1 and the path of the steepest descent Γ_0. However, when the path of steepest descent Γ_0 lies in close proximity of the pole, special precautions must be taken in the evaluation of the integral of (30). This is illustrated in Appendix B. The pole P_1 is of no concern since it lies on the second Riemann sheet of the branch point β_{B_2} on which $\operatorname{Im}\left(\sqrt{\varepsilon-\sin^2\beta}\right) > 0$. The presence of the branch point B_2^3 should ordinarily be taken into account when ε is close to unity and when deforming the path Γ_1 to the path of steepest descent Γ_0. Often, the contribution along the borders of the branch cut would be a fast decreasing exponential that can be neglected in comparison to the contribution from the saddle point integration. In short, (30) becomes

$$g_{sV} \approx \int_{\Gamma_0} e^{-j\pi/4} \sqrt{\frac{2k_1\sin\beta}{\pi R_2\sin\theta}} \frac{\sqrt{\varepsilon-\sin^2\beta}}{\varepsilon\cos\beta+\sqrt{\varepsilon-\sin^2\beta}} \exp\left[-jk_1R_2\cos\left(\beta-\theta\right)\right]d\beta \tag{34}$$

where Γ_0 is a path in the complex β plane as shown in Figure 4. The path of steepest descent never crosses any of the poles. The contributions along the borders of the second branch cut associated with the branch point k_2, particularly for low values of the dielectric constant ε are not necessary as they would be fast decreasing exponentials [19-27] that can be neglected in comparison to the contribution from the saddle point integration. Hence, by application of the method of steepest descent to (34), one obtains for $\theta < \pi/2$, (i.e., when the pole is not near the saddle point) applying the methodology described in Appendix B, one obtains.

$$g_{sV} \approx \frac{2\exp\left[-jk_1R_2\right]}{R_2} \frac{\sqrt{\varepsilon-\sin^2\theta}}{\varepsilon\cos\theta+\sqrt{\varepsilon-\sin^2\theta}} \left[1-\frac{1}{2jk_1R_2}\times \left\{ \frac{\varepsilon(\varepsilon-1)\left[2\varepsilon(\varepsilon-1)+\varepsilon\cos^2\theta\left(3-\cos^2\theta\right)+\cos\theta\sqrt{\varepsilon-\sin^2\theta}\left(2\varepsilon+\sin^2\theta\right)\right]}{\left(\varepsilon-\sin^2\theta\right)^2\left[\varepsilon\cos\theta+\sqrt{\varepsilon-\sin^2\theta}\right]^2} - \frac{1}{4\sin^2\theta} \right\}\right] \tag{35}$$

Hence $\Pi_{1z}^{reflected}$ of (16) can now be written as

$$\Pi_{1z}^{reflected} \approx P\frac{\exp\left(-j\,k_1\,R_2\right)}{R_2}\left[\frac{\varepsilon\cos\,\theta-\sqrt{\varepsilon-\sin^2\theta}}{\varepsilon\cos\,\theta+\sqrt{\varepsilon-\sin^2\theta}}+\frac{1}{jk_1R_2}\times\right. \tag{36}$$

$$\left.\left\{\frac{\varepsilon\left(\varepsilon-1\right)\left[2\varepsilon\left(\varepsilon-1\right)+\varepsilon\cos^2\theta\left(3-\cos^2\theta\right)+\cos\,\theta\left(2\varepsilon+\sin^2\theta\right)\sqrt{\varepsilon-\sin^2\theta}\right]}{\left(\varepsilon-\sin^2\theta\right)^{3/2}\left[\varepsilon\cos\,\theta+\sqrt{\varepsilon-\sin^2\theta}\right]^3}-\frac{\sqrt{\varepsilon-\sin^2\theta}}{2\sin^2\theta\left[\varepsilon\cos\theta+\sqrt{\varepsilon-\sin^2\theta}\right]}\right\}\right]$$

The first term of (36) represents a spherical wave originating from the image and can be rewritten as

$$\Pi_{1z}^{reflected} \approx P\,\Gamma_{TM}\frac{\exp\left(-j\,k_1R_2\right)}{R_2} \tag{37}$$

where Γ_{TM} can be recognized *as the TM reflection coefficient associated with the spherical wave* [3,15-26], and is given by

$$\Gamma_{TM}=\frac{\varepsilon\cos\,\theta-\sqrt{\varepsilon-\sin^2\theta}}{\varepsilon\cos\,\theta+\sqrt{\varepsilon-\sin^2\theta}} \tag{38}$$

The name reflection coefficient method is derived from (37) since $\Pi_{1z}^{reflected}$ is now obtained as the reflection coefficient times the potential from the image of the source. **The method represents a good approximation, as long as the fields are computed far away from the ground plane and away also from the source dipole to ensure** $\theta<\pi/2$. This implies that the use of the reflection coefficient in the computations of the reflected fields are **NOT VALID NEAR THE GROUND**, where $\theta\approx\pi/2$ [19,25]. The antennas (both transmit and receive) need to be elevated some distance from the ground for the reflection coefficient method to be mathematically meaningful. Hence the two ray model generally used in wireless propagation analysis may thus yield questionable results! The total Hertz potential in medium 1, when the conductivity of the relative permittivity of the lower medium is large, i.e., $|\varepsilon|>1$, is given by

$$\Pi_{1z}\approx P\left[\frac{\exp\left(-jk_1R_1\right)}{R_1}+\frac{\exp\left(-jk_1R_2\right)}{R_2}\left\{\frac{\sqrt{\varepsilon}\cos\,\theta-1}{\sqrt{\varepsilon}\cos\,\theta+1}+\frac{2\varepsilon}{j\,k_1R_2}\left(\frac{1}{\sqrt{\varepsilon}\cos\theta+1}\right)^3+...\right\}\right] \tag{39}$$

Note that when $|\varepsilon|\to\infty$, Π_{1z} of (39) goes properly into the form of a source plus an image term due to a vertical electric dipole located above a perfectly conducting ground plane. However when $\theta\approx\pi/2$, and $\sqrt{\varepsilon}\cos\theta$ □ 1 it becomes

$$\Pi_{1z}\approx P\left[\frac{\exp\left(-jk_1R_1\right)}{R_1}-\frac{\exp\left(-jk_1R_2\right)}{R_2}+\frac{2\varepsilon}{jk_1R_2^2}\exp\left(-jk_1R_2\right)+...\right] \tag{40}$$

It is now important to recognize from (37) that the sum of the first two terms may be smaller than the third term. As a matter of fact when both the transmitter and the receiver are near the ground, i.e.,

$$R_1=R_2\approx\rho \quad ;\text{ and }\quad z\approx 0\approx z' \tag{41}$$

then observe that the fields will solely be determined by the third and higher order terms of (40). This may result in poor convergence as the sum of all the three terms are dictated by the higher order terms of the saddle point integration. The reason for this poor convergence in the vicinity of $\theta\approx{}^{\pi}\!/_{2}$ is that the effect of the pole close to ${}^{\pi}\!/_{2}$ becomes important. The bottom line is since it is the higher order terms that are responsible for the calculation of the fields along the interface in (40), we need to carry out a different asymptotic expansion as illustrated in Appendix C. The conventional saddle point method cannot now be applied as there is a pole near a saddle point. This necessitates the development of an alternate saddle point method. The treatment of a pole near a saddle point can be handled in two different ways. The singularity of the pole can be handled either in an additive fashion as presented by Collin [3] and Tyras [26] or in a multiplicative form as has been discussed by Clemmow [32] or Wait [20]. Details of the procedure can be found in Felsen and Marcuvitz [33]. However, when factored into the total solution they yield similar results. Here we use the multiplicative form of the singularity as it provides a simpler methodology.

The mathematical details are included in the appendix C for completeness of the presentation. If one applies the modified saddle point method of evaluating the integral of the Green's function given by (34) using the methodology of Appendix C where the singular part of the integral is factored into

$$\frac{1}{\left[\varepsilon\cos\beta+\sqrt{\varepsilon-\sin^2\beta}\right]}=\frac{1}{\varepsilon^2-1}\times\frac{\sqrt{\varepsilon-\sin^2\beta}-\varepsilon\cos\beta}{\sin\left(\beta+\beta_P\right)\sin\left(\beta-\beta_P\right)} \tag{42}$$

where $\varepsilon\cos\beta_P+\sqrt{\varepsilon-\sin^2\beta_P}=0$ with $\sin\beta_P=\pm\sqrt{\frac{\varepsilon}{\varepsilon+1}}$ and $\cos\beta_P=-\sqrt{\frac{1}{\varepsilon+1}}$. Applying the modified steepest descent methodology where a pole may be located near the saddle point path of integration results in (as per Appendix C)

$$\begin{aligned} g_{sV}&\approx\exp\left(\frac{-j\pi}{4}\right)\sqrt{\frac{2k_1}{\pi R_2}}\frac{(\sqrt{\varepsilon-\sin^2\theta})\left(\sqrt{\varepsilon-\sin^2\theta}-\varepsilon\cos\theta\right)}{(\varepsilon^2-1)(\cos\theta-1./\sqrt{\varepsilon+1})(2\sin\frac{(\theta+\beta_P)}{2})}\int_{\Gamma_0}\frac{\exp\left[-jk_1R_2\cos\left(\beta-\theta\right)\right]}{\sin\frac{(\beta-\beta_P)}{2}}d\beta\\ &\approx\sqrt{\frac{4\pi k_1 j}{R_2}}\;\frac{\sqrt{\varepsilon-\sin^2\theta}}{\varepsilon^2-1}\;\frac{\sqrt{\varepsilon-\sin^2\theta}-\varepsilon\cos\theta}{\cos\theta-1./\sqrt{\varepsilon+1}}\;\frac{\exp\left[-jk_1R_2-W^2\right]erfc(jW)}{\sqrt{1+\frac{\cos\theta}{\sqrt{\varepsilon+1}}+\frac{\sqrt{\varepsilon}\sin\theta}{\sqrt{\varepsilon+1}}}}\end{aligned} \tag{43}$$

with W - t*he numerical distance* - termed by Sommerfeld [19,34] is given by

$$W^2=-j\,k_1R_2 2\sin^2\left(\frac{\theta-\beta_P}{2}\right)=-jk_1R_2\left[1+\frac{\cos\,\theta}{\sqrt{\varepsilon+1}}-\frac{\sqrt{\varepsilon}\sin\theta}{\sqrt{\varepsilon+1}}\right] \tag{44}$$

Upto this point there is no controversy at all and all the formulations including the original one developed by Sommerfeld agree with each other. The disagreements show up when various approximations are made of the term $\exp\left[-W^2\right]erfc(jW)$ **for different assumptions.**

For example, if we require | ε | >> 1, and for close to normal incidence

$$\frac{\sqrt{\varepsilon-\sin^2\theta}}{\varepsilon^2-1}\approx\frac{\sqrt{\varepsilon}}{\varepsilon^2-1}\approx\frac{1}{\varepsilon^{1.5}}\;\;;\;\;\frac{\sqrt{\varepsilon-\sin^2\theta}-\varepsilon\cos\theta}{\cos\theta-1./\sqrt{\varepsilon+1}}\approx-\varepsilon\;;\;\sqrt{1+\frac{\cos\theta}{\sqrt{\varepsilon+1}}+\frac{\sqrt{\varepsilon}\sin\theta}{\sqrt{\varepsilon+1}}}\approx 1 \tag{45}$$

If we further restrict that the observation point is close to the source, then

$$\exp\left[-jk_1R_2-W^2\right]erfc(jW)\approx 1\;\text{ for small values of } W \tag{46}$$

After incorporating all these approximations, one obtains

$$g_{sV} \approx -\sqrt{\frac{2\pi k_1 j}{\varepsilon}}\,\frac{\exp\left[-jk_1R_2\right]}{\sqrt{R_2}} \tag{47}$$

By incorporating (47) into (20) results in

$$\Pi_{1z} = P\left[g_0 + g_1 - g_{sV}\right] = P\left[\frac{\exp\left(-jk_1R_1\right)}{R_1} + \frac{\exp\left(-jk_1R_2\right)}{R_2} + \sqrt{\frac{2\pi k_1 j}{\varepsilon}}\,\frac{\exp\left[-jk_1R_2\right]}{\sqrt{R_2}}\right] \tag{48}$$

Observe that the elusive controversial surface wave terms shows up when one uses these particular approximation mentioned earlier for the various parameters including the complementary error function! Hence there was no error in sign in the original Sommerfeld formulation. And it is a myth that was propagated by researchers who probably could not scrutinize in details Sommerfeld's derivation as no mention has ever been made as to where Sommerfeld made the error in his derivations.

It is the expression of (48) that made Sommerfeld to comment: *While at short distances from the transmitter, the waves are almost entirely of the nature of space waves, as the distance increases the surface component becomes more and more predominant, as its amplitude decreases more slowly than that of the surface component. That is, the nature of the wave constantly approaches that of a surface wave. When the distance becomes very great, the surface wave may again give way to the space wave, as the former is more rapidly absorbed. It is questionable, however, whether this effect is of practical importance.* It is these statements made by Sommerfeld that led to a controversy and contributed to subsequent confusion and the erroneous sign error mythology in his original paper!

To compute the fields near the interface we use not the expression of (20) discussed so far but that of (21). This is because near the interface when $\theta \approx \pi/2$ the TM reflection coefficient of (19) becomes -1 and so the expressions of (21) are more appropriate than that of (20). One possible reason for using (21) is given by Stratton [20] as the reflection coefficient is approximately +1 for a perfectly conducting ground when the fields are observed far from the ground and it transforms to -1 when the fields are observed near the ground when, $\theta \approx \pi/2$. Hence a different expansion of the Hertz potential is necessary as it will provide a more accurate approximation. To this effect the form (21) was chosen over (20).

In order to solve for the total fields near the interface [3], a modified saddle point method as explained in Appendix C [19,54] is applied to take into account the effect of the pole β_P near the saddle point. In the expression of G_{SV} in (31) there is a pole β_P which is seen from [19,54] and the final solution is given by

$$\begin{aligned} G_{sV} &= \varepsilon \exp\left(-j\pi/4\right) \int_{\Gamma_1} \left(\frac{2k_1\sin\beta}{\pi R_2\sin\theta}\right)^{1/2} \frac{\exp\left[-jk_1R_2\cos\left(\beta-\theta\right)\right]\cos\beta}{\varepsilon\cos\beta + \sqrt{\varepsilon - \sin^2\beta}}\,d\beta \\ &\approx \varepsilon\sqrt{\frac{4\pi k_1 j}{R_2}}\,\frac{\cos\theta}{\cos\theta - \frac{1}{\sqrt{\varepsilon+1}}}\,\frac{\sqrt{\varepsilon - \sin^2\theta} - \varepsilon\cos\theta}{\varepsilon^2 - 1}\,\frac{\exp\left[-jk_1R_2 - W^2\right]erfc\left(jW\right)}{\sqrt{1 + \frac{\cos\theta}{\sqrt{\varepsilon+1}} + \frac{\sqrt{\varepsilon}\sin\theta}{\sqrt{\varepsilon+1}}}} \end{aligned} \tag{49}$$

where W is given by (45).

Now if one considers, the case, where $|\varepsilon| > 1$ and $\theta \approx \pi/2$ and W is very small then we have $\exp\left[-W^2\right]erfc\left(jW\right) \approx 1$. Under this assumption, one obtains the following

$$\frac{\sqrt{\varepsilon-\sin^2\theta}-\varepsilon\cos\theta}{\varepsilon^2-1}\approx-\frac{1}{\varepsilon}\quad;\quad\frac{\cos\theta}{\cos\theta-1./\sqrt{\varepsilon+1}}\approx 1\;;\quad\sqrt{1+\frac{\cos\theta}{\sqrt{\varepsilon+1}}+\frac{\sqrt{\varepsilon}\sin\theta}{\sqrt{\varepsilon+1}}}\approx 1 \tag{50}$$

and

$$G_{sV}\approx-\sqrt{\frac{4\pi k_1 j}{R_2}}\exp\left[-jk_1R_2\right] \tag{51}$$

Again, the controversial surface wave term shows up! However, the interpretations are made using lots of approximations some of which may be questionable and the bottom line is one should look at the total solution and not the component ones.

Now if one considers, the second case, where $|\varepsilon| > 1$ and $\theta \approx \pi / 2$ and W is very small and (46) is satisfied then we have

$$G_{sV}\approx-\sqrt{\frac{2\pi k_1 j}{R_2}}\exp\left[-jk_1R_2\right]\frac{(z+z')}{R_2}\frac{\varepsilon}{\sqrt{\varepsilon^2-1}}\approx-\sqrt{2\pi k_1 j}\;\frac{(z+z')\exp\left[-jk_1R_2\right]}{R_2^{1.5}}. \tag{52}$$

Equation (52) thus illustrates that when $\theta\approx\pi/2$ the dominant term of the potential $\Pi_{1z}\propto\frac{1}{R_2^{1\cdot5}}$ and therefore the leading term for the fields will be also be varying as $\frac{1}{\rho^{1\cdot5}}$, if $(z + z')$ is small compared to ρ in (52). It is interesting to observe that Eq. (52) is not a function of the ground parameters. So the path loss exponent factor in a mobile urban cellular communication should be 3 near the ground and the reflection coefficient method is not applicable, under those circumstances. This should approximately hold for any types of ground parameters, like urban, or suburban or even lake and oceans [19]. As we shall see in the next section, this component of the solution has the characteristics of a lateral wave.

However as W becomes large then

$$\exp\left[-W^2\right]erfc\left(jW\right)\approx\frac{-j}{W\sqrt{\pi}}\left[1+\frac{1}{2W^2}\right] \tag{53}$$

and for $|\varepsilon|>1$,

$$W^2\approx\frac{-j\,k_1\,R_2}{2\,\varepsilon}. \tag{54}$$

Under this condition,

$$G_{sV}\approx 2\sqrt{\varepsilon}\exp\left[-jk_1R_2\right]\frac{(z+z')}{R_2^2}\left[1-\frac{\varepsilon}{j\,k_1R_2}\right] \tag{55}$$

Thus the total Hertz potential in medium 1 which is valid near the interface for $|\varepsilon|>1$ and $\theta\approx\pi/2$ becomes:

$$\Pi_{1z}\approx\begin{cases}P\left[\dfrac{\exp(-jk_1R_1)}{R_1}-\dfrac{\exp(-jk_1R_2)}{R_2}-\sqrt{j2\pi k_1}\,(z+z')\dfrac{\exp(-jk_1R_2)}{R_2^{1.5}}\right] & ,\,W<1\\ P\left[\dfrac{\exp(-jk_1R_1)}{R_1}-\dfrac{\exp(-jk_1R_2)}{R_2}+2\sqrt{\varepsilon}(z+z')\dfrac{\exp(-jk_1R_2)}{R_2^{\,2}}\left[1-\dfrac{\varepsilon}{jk_1R_2}\right]\right] & ,\,W>1\end{cases} \tag{56}$$

Indeed (56) is exactly what Okumura et. al experimental data illustrates as he presented in his classic paper [19,36]

The above simplified expressions illustrate that a ground wave [34] decays asymptotically as $1/R^2$ and this applies only in the far field region, where $W > 1$, as the first two terms cancel in the second expression. Also, it is interesting to note that the third term for $W > 1$ provides the so called height-gain for the transmitting and receiving antennas. However, this height gain applies to both intermediate and far field regions. In the intermediate region, the fields decay as approximately $\rho^{-1.5}$. Also, observe that for $W < 1$, the above expression is independent of the ground parameters. This is confirmed in [34], using a more accurate numerical analysis. It is important to note that this Sommerfeld representation for the fields is not valid when z and z' are close to 0, as can be seen from (7) and (17).

In summary, there is no controversy whatsoever in the mathematical expressions derived so far related to the Green's functions and after they have been evaluated using the modified saddle point of integration. The real cause of the source of disagreements between various authors that Sommerfeld made an error in the sign in his 1909 paper has no basis as the error in the sign is a myth. The only correct and clear explanation in our opinion of the origin of this mythology and discrepancy is given by Collin [3]: *The end of the story is that Sommerfeld's solution had an error in sign has no merit. Sommerfeld's first solution is given by his asymptotic series plus the Zenneck surface wave. His second solution is given by a power series, which is consistent with his first solution. It is Sommerfeld's asymptotic series and power-series expansions that provide the clues for identifying the flaws in the claims that Sommerfeld had made a sign error. There are inherent limitations in Sommerfeld's solution, but they are not caused by a sign error*."

3. EPILOGUE TO THE SOMMERFELD SIGN ERROR SAGA

In short, Sommerfeld in 1909 [16] undertook a detailed analysis of the radiation from an infinitesimal vertical Hertzian dipole over an imperfectly conducting infinite ground medium to complete the demonstration of the surface wave component of the total field. Sommerfeld obtained the phasor for the potential from a dipole located in air (medium 1 with a propagation constant k_1) and radiating over a ground plane having a complex permittivity ε (medium 2 with a propagation constant $k_2 = k_1\sqrt{\varepsilon}$). The denominator for the expression of the fields had branch points at $\lambda = \pm k_1$ and $\lambda = \pm k_2$, and poles at $\lambda_P = \pm\frac{k_1 k_2}{\sqrt{k_1^2 + k_2^2}}$. The corresponding Riemann surface has four sheets; on only one of these are fulfilled the necessary conditions for the convergence of the integral at infinity. According to Sommerfeld, the path of integration can be resolved on this sheet in three parts: the first one is a loop from infinity about the branch point $\lambda = k_1$, the second one is a similar loop about $\lambda = k_2$ and the third one could be any small circle about the pole $\lambda = \lambda_P$ as seen in Figure 3. The contributions of the loops about the branch points give the dominant terms at $\frac{e^{-jk_1R}}{R}$ and $\frac{e^{-jk_2R}}{R}$ which can be identified as the space waves while the residue at the pole has a variation of the form $\frac{e^{-j\lambda_P R}}{\sqrt{R}}$ which had all the hall marks of a true surface wave. Here, R is the distance between the source and the observation (field) point.

As part of his solution, Sommerfeld illustrated that a surface wave contribution arose from the pole of the integrand (i.e., from the solution for λ in the following equation $\left[k_2^2\sqrt{\lambda^2 - k_1^2} + k_1^2\sqrt{\lambda^2 - k_2^2} \;= 0\right]$), and other radiated waves evolved from the branch cut contributions related to the two branch points located at k_1 and k_2. It was recognized from the onset by Sommerfeld that his total solution of (1) could be interpreted as a bundle of plane waves reflected and refracted from the surface of the earth at various angles of incidence. The surface integrals are extended over the plane earth and over small spheres which exclude the singularities occurring at the source and at the point of observation. Sommerfeld's approach was based on a deformation of the path of integration in the complex λ - plane as shown in Figure 3.

In short, historically, in 1909 Sommerfeld computed the integral along the positive real axis of Figure 3 by first applying the Cauchy principal integral method to close the contour by a large semicircle at infinity given by the semi- circular contour with indentations of Figure 3a lying in the third and the fourth quadrants. The resultant is the integral given by the contours of Figure 3a. As seen in his book [22], this closed contour of integration is equivalent to two integrals around the branch cuts associated with the branch points at $+k_1$ and $+k_2$ and a contour integration around the pole $\lambda = \lambda_P$ where $\lambda_P = \dfrac{k_1 k_2}{\sqrt{k_1^2 + k_2^2}}$. The other branch points $-k_1$ and $-k_2$ and the pole located at $-\dfrac{k_1 k_2}{\sqrt{k_1^2 + k_2^2}}$ are of no concern as they are located in the upper half plane where the contour is not closed as seen in Figure 3. Sommerfeld then evaluated the residue at the pole and showed that it has the form of a *surface wave*. Sommerfeld came to the conclusion from

$$\Pi_{1z}^{pole} = -2\pi j P \left[\frac{k_2^2 H_0^{(2)}(\lambda_P \rho) \exp\left(-z\sqrt{\lambda_P{}^2 - k_1{}^2}\right)}{\dfrac{k_2^2}{\sqrt{\lambda_P{}^2 - k_1{}^2}} + \dfrac{k_1^2}{\sqrt{\lambda_P{}^2 - k_2{}^2}}} \right] \tag{57}$$

where Π_{1z}^{pole} is part of the solution from the pole contribution. For large values of ρ, the asymptotic representation for the Hankel function is used

$$\mathrm{H}_0^{(2)}(\lambda\rho) \xrightarrow[|\lambda\rho| \to \infty]{\text{Limit}} \sqrt{\frac{2}{\pi\lambda\rho}} \exp\left[-j\lambda\rho + j\pi/4\right] \tag{58}$$

resulting in

$$\Pi_{1z}^{pole} = P \left[\sqrt{\frac{2\pi}{j\lambda_P \rho}} \frac{k_2^2 \exp\left(-j\lambda_P \rho - z\sqrt{\lambda_P{}^2 - k_1{}^2}\right)}{\dfrac{k_2^2}{\sqrt{\lambda_P{}^2 - k_1{}^2}} + \dfrac{k_1^2}{\sqrt{\lambda_P{}^2 - k_2{}^2}}} \right] \tag{59}$$

As Sommerfeld in his book then points out: *this formula bear all the marks of surface waves.* [a true surface wave is a slow wave and the fields become concentrated to the interface as the frequency increases]. "*It was the main point of the author's work of 1909 to show that the surface wave fields are automatically contained in the wave complex. This fact has of course, not changed. What has changed is the weight which we attached to it. At that time it seemed conceivable to explain the overcoming of the earth's curvature by radio signals with the help of the character of the surface waves; however we know now that it is due to the ionosphere. In any case, the recurrent discussion in the literature on the reality of the Zenneck waves seems immaterial to us*."

Thus, the initial results Sommerfeld obtained lent considerable credence to Marconi's view that the electromagnetic wave was guided along the surface. As Collin noted in [3], as early as 1902, Kennelly [37] and Heaviside [38] predicted the existence of an ionized layer at considerable height above the surface of the earth. It was thought that such a layer could possibly reflect the electromagnetic waves back to earth and it was experimentally verified by Brett and Tuve in 1926 [39]. So no serious challenge to the surface-wave mechanism for long distance propagation occurred until a decade later when Weyl published a paper on the same subject and obtained a solution very similar to that found by Sommerfeld [16], but without the surface wave term [40]. Weyl [40] obtained an asymptotic series representing the diffracted field by applying a method of steepest descent. Weyl's solution also reduces to a form which can be interpreted as the superposition of a space and a surface wave, but the Weyl surface wave is not identical to that of Sommerfeld [16] and Zenneck [15].

In 1926, Sommerfeld returned to the same problem, and this time solved it [17] using a different approach, and confirmed the correctness of Weyl's solution [40] "*which he says is best represented by the one contour integral that goes near the pole and the second saddle point at k_2*" [17]. . With a better understanding of the ionospheric mode of propagation, the concept of the surface wave being the important factor for long-distance propagation lost favor. It is the presence of the Kennelly-Heaviside layer of the ionosphere that is responsible for the long-distance propagation of the long wavelength fields. At this point, it is important to note that Sommerfeld never refer to an error in the sign in his original work.

In 1930, Van der Pol and Niessen published a new solution to the old Sommerfeld problem, using yet another method of solution [40]. Again the results of Weyl and the later results of Sommerfeld were confirmed. This was followed by another paper of Van der Pol on the same problem [41]. Each independent solution of the old Sommerfeld problem agreed with his 1909 solution except the surface wave term.

As Collin points out [3]: *Sommerfeld's solution as claimed by many that there is an error in sign has no merit. Sommerfeld's first solution is given by his asymptotic series plus the Zenneck surface wave. His second solution is given by a power series, which is consistent with his first solution. It is Sommerfeld's asymptotic series and power-series expansions that provide the clues for identifying the flaws in the claims that Sommerfeld had made a sign error. There are inherent limitations in Sommerfeld's solution, but they are not caused by a sign error.* This is exactly what we illustrate in this paper that different types of approximations for the Hertz potentials yield different expressions and all of them are correct even though they look seemingly different.

In the 1930s, Norton undertook the task of reducing the formulas of Van der Pol and Niessen to practical form for the radio engineer [43-45]. As a part of this undertaking, he apparently believed that he had found a sign error in Sommerfeld's 1909 paper. In 1935, Norton published a short paper in which he asserted that Sommerfeld had made an error in sign in one of his formulas [46]. Unfortunately, Norton did not provide any specific details as to which of Sommerfeld's expressions had the sign error, or what had gone wrong in Sommerfeld's analysis. In 1937, Niessen [47] published a paper in which he also claimed that Sommerfeld had made a sign error in his 1909 paper. According to Niessen, the sign error came about because Sommerfeld did not take the value of the argument of the square root of a complex number using the convention that this should always be taken to be between 0 and 2π .

An intensive analysis was carried out in AT&T Bell Labs regarding Sommerfeld's original formulation but no error was found! This led Charles Burrows of the Bell Laboratories to carry out experiments. Charles Burrows [48,49], carefully measured the field strength at 150 MHz, for a distance ranging from 1 to 2000 m over a deep, calm, lake near Seneca in upper New York. He showed that his measured results did not support a surface wave. His experimental results were in conformity with Weyl-Norton's theory, namely, in the far field a surface wave type phenomenon is not observed.

He further concludes that these results along with the experiments proved that simple antennas do not generate a Sommerfeld surface wave [49].

A conference was held at Bell Labs in 1935 to determine which theory was correct, and the conference attendees were faced with a dilemma in that they could find no error in either Sommerfeld's or Weyl's analysis. After a half-day conference …was unsuccessful in finding the source of any error in either paper [Sommerfeld (1909) or Weyl (1919)], Dr. Fry suggested the experimental approach."[50].

In addition Wise [51] published a paper which showed that a vertical dipole does not generate a surface wave which at great distances behaves like Zenneck's plane wave or a surface wave. A contemporary theoretical investigation by S. O. Rice [52] lead to the same conclusion.

As summarized by Stratton [21, p.583] that the reflection coefficient has four sheets and on only one of these are the conditions fulfilled necessary for the convergence of the integral at infinity. According to Sommerfeld the path of integration can be resolved on this sheet into three parts: the first a loop from infinity about the branch point $+k_1$., the second a similar loop about $+k_2$, and the third any small circle about the pole at λ_p. The contributions of the loops about the branch points Q_1 and Q_2 while the residue at the pole gives a term P, so that $\Pi_1 = Q_1 + Q_2 + P$. The terms Q_1 and Q_2 being proportional to $\frac{\exp(-j\,k_1\,R)}{R}$ and $\frac{\exp(-j\,k_2\,R)}{R}$. On the other hand, for P Sommerfeld obtained a function of the form of the surface wave. Asymptotic expansions have been given by Wise and by Rice which show that the term P in Sommerfeld's solution is canceled when all the terms of the series for Q_1 , Q_2 and P are taken into account. This was independently verified by Collin [3].

In spite of the apparent closure of the analytical debate, the controversies would not die. The late Kenneth Norton, an eminent radio engineer in the USA, exchanged with Sommerfeld numerous letters on this topic. An example is a letter [53], from Sommerfeld to Norton, which was written while the former was on holiday in the Austrian Tyrol. Sommerfeld never agreed that an error or miscue had ever been made. Two later letters from Norton [54,55] indicate his views on the separation of surface waves and space waves. Sommerfeld acknowledged Norton's communications and suggested he compare his results with those of Van der Pol and Niessen [41]. Of course, there is a consistency here, because Norton [43-45] based his papers on the results of Van der Pol and Niessen [41]!

As stated by Wait [2], the confusion was also caused when Sommerfeld in 1926 [17] published a paper where he revisited his old paper of 1909 and changed the sign of the argument of the complex error function without any proper explanations. Subsequently, there have been many speculations that Sommerfeld in the 1909 paper had an error in the sign which he corrected in the 1926 paper. Since Weyl's method seemed mathematically simpler his result was favored by public opinion in numerous papers by other authors. Finally in 1935, Sommerfeld himself conceded that the surface wave has no reality. But he never admitted to an error in the sign in his 1909 paper! Referring to F. Noether [54], he attributed this to an inaccuracy in the evaluation of his general solution. According to Noether's explanation the pole is so close to one of the branch cuts that the integration method used by Sommerfeld possibly was not sufficiently reliable.

Baños [25] states that despite Ott's [57] valuable contribution on the question of the existence or non-existence of Sommerfeld's electromagnetic surface wave, this topic continued to be debated in the literature. In 1947 Epstein [58] published a paper in which he disqualified Sommerfeld's original formulation of the problem by proposing a new contour of integration excluding the pole which was shown to be in error by Bouwkamp [57]. Next Kahan and Eckart [60,61] published a series of papers and got their acts right in the latter ones where they claimed, according to Bouwkamp [59], to accept Sommerfeld's original solution and pointed out that Sommerfeld's evaluation of the integral around the branch cut was in error. They stress that the correct evaluation would have yielded an expression that contains the surface wave with negative sign, so that the real result would have coincided with Weyl's result, the negative surface wave term being cancelled by the positive term due to the residue of the pole of the integrand. This explanation and clarification of the controversy is similar to that of Wise [49] and Rice [50] who showed that the existence of a surface wave is not a part of the total solution. Also Boukamp [59] had doubts about the accuracy of the analysis of the previous authors.

The final statements of this saga can be concluded by a quote from Goubau. The very elegant remark made by Goubau in his IRE Transactions paper [62] on *Waves on interfaces* in 1959 is quite revealing in clarifying the situation: "*The existence of the Zenneck wave has been disputed so much in the literature that it may be excusable if a few more remarks are added to this subject. The dispute originated in the attempt to extract from the mathematical solution of a problem more answers than there were questions when the problem was formulated mathematically.*

If one formulates the problem of a dipole radiating above a plane interface, one can only expect a solution which describes the total field of the dipole, no matter what mathematical method one uses. There is no obligation from the physical point of view to use complex variables for solving the problem. The fact that the problem is more accessible to a rigorous treatment, when solved in the complex plane, is a purely mathematical matter and one cannot expect this method to yield more information than a treatment with only real quantities. If Sommerfeld had solved the problem by use of real quantities only, it is unlikely that the question as to a surface wave and a space wave would ever have arisen (except perhaps now) since the actual field of a dipole at the interface differs substantially from that of the surface wave.

The fact that the answer to a given problem can be written in terms of a complex integral whose integrand comprises a pole, does not a priori mean that this pole has physical meaning. Even if this is the case, there is still no obligation that the integration has to be performed in such a manner that the pole is included.

In order to separate the field into two components, it is necessary to define both these components. Defining only one, namely the surface wave, is not enough. However, I believe that the orthogonality relations quoted in this paper should fill this gap and present a satisfactory basis for the separation".

In summary, the real cause of the source of disagreements between various authors that Sommerfeld made an error in the sign in his 1909 paper has no basis as the error in the sign is a myth. Since the issue is on the proper interpretation of the mathematical results an overview of the various types of waves are discussed and it is also illustrated as how they are interpreted based on their pole locations with respect to the assumed contours of the branch cut. The important point is that some of the definitions of the various waves are quite arbitrary as is illustrated next.

4. ILLUSTRATION OF THE VARIOUS WAVES INCLUDING ZENNECK, SURFACE, LATERAL AND LEAKY WAVES

There are various terms [63,64] that are used to describe a solution to Maxwell's equations. They are either characterized by proper modes, quasimodes, surface waves, leaky waves, lateral waves and radiation fields from the structure. The origin of these terminologies and their various interpretations are now looked at. The reason is that the terminologies are a function of the location of the pole with respect to the Riemann sheets associated with the saddle point integration path for the Green's function. The final results are then determined by the nature of the assumed branch cuts associated with the various branch points and their appropriate locations which will be explained later. All these terminologies are needed in the complete description of the field, and their relative intensities in describing a complete solution. As Marcuvitz [63] pointed out: *Solutions to source-excited field problems are frequently represented as superposition of source-free field solutions. The latter are in general of two types: eigenmodes and noneigen-modes which are related to the zeros of the total impedance or alternatively the poles of the scattering coefficient of a system. The eigenmodes are* *everywhere finite and comprise a complete orthogonal set. The noneigenmodes become infinite in the infinitely remote spatial limits of a region and are not in general members of a complete orthogonal set; examples are "radio-active states," "damped resonances," and "leaky waves." Despite their physically singular behavior, the nonmodal solutions can be employed to represent field solutions in certain ranges. In regions of finite extent bounded by impermeable walls; i.e., "closed regions," the source-free solutions generally possess orthogonality and completeness properties that permit an arbitrary function, and in particular a desired field solution, to be represented by their superposition. Such source-free solutions are termed the characteristic (eigen) or normal modes of the given region; they possess a discrete spectrum, are everywhere finite, and individually satisfy the field equations plus appropriate boundary conditions. In regions of infinite extent; i.e., "open regions," there may exist a corresponding discrete spectrum of modes but for completeness these must be supplemented in general by a continuous spectrum of characteristic modes to permit the representation of an arbitrary function. In contrast to the discrete modes, the continuous modes are improper in that individually they are not absolute square integrable (with finite energy) nor do they satisfy the requisite boundary conditions at the singular point, infinity. In the presence of a continuous spectrum there may also exist a set of nonmodal solutions of the source-free field equations which in distinction to the modes become infinite in the infinitely remote limits of an open region; in this category belong the so-called "radio-active state," "damped resonance," and "leaky wave" solutions. Despite their physically unacceptable behavior in the given region, such nonmodal solutions may nevertheless be employed for the representation of field solutions in an extended region of which the given region is a subspace. Although not in general members of a complete set of orthgonal functions, the nonmodal solutions may nevertheless be employed to obtain rapidly convergent field representations in certain ranges. A pragmatic characterization of both the nonmodal and the discrete modal solutions stems from their connection with the zeros of an impedance of a system. As is well known, the zeros of the total impedance (admittance) or, alternatively, the poles of the scattering coefficient of a system represent resonant frequencies that distinguish all possible source-free solutions. For a conservative (hermitian) system real resonant frequencies represent eigenmodes whereas complex resonant frequencies characterize the nonmodal solutions. In the case of nonconservative; i.e., dissipative, systems both the modal and nonmodal frequencies are in general complex and one of the points to be discussed is how they are to be distinguished.*

From the earliest days in the history of electromagnetic theory it has been customary to draw distinction between “free waves” and “bound” or “guided waves.” Yet, from among the great wealth of the written information, it is impossible to extract a clear distinction between the two types of waves. Free waves are thought to be electromagnetic waves which spread out, from a source, in a medium of infinite extent. Guided waves, on the other hand, are associated with boundaries which confine or channel the electromagnetic energy in accordance with the properties and the manner in which the boundaries are distributed. Such a distinction, though adequate in many cases, is confusing in others. However, confusion still exists.

As stated by Karbowiak [64]: *The early work of Sommerfeld is a good illustration of the problems involved. In its original formulation, the problem is to find an expression for the field at a long distance from a Hertzian dipole situated at a certain height above a plane earth. The formal solution to the problem (in the form of a contour integral) presents no fundamental difficulty and the integral expression obtained is, mathematically, certainly correct; the difficulty lies elsewhere. The closed form, a rigorous solution in the form of an integral, is devoid of physical meaning and is incapable of useful and direct physical interpretation. A more difficult part of the problem is now the interpretation of the closed-form solution, using suitable approximations, and it is precisely from this point onwards* *that the opinions, methods of approach and interpretation of the results differ, as obtained by various investigators. Sommerfeld himself obtained approximate solutions by deforming the path of integration and expanding the resulting integral in a suitable asymptotic series. The solution was composed of a "space wave" and a "surface wave." whereas all concerned seem to have understood the meaning and physical interpretation of "space wave" (radiation), "surface wave" became a topic of lengthy discussions for decades. More recently solutions have been obtained to a number of allied problems all concerned with the field produced by a dipole (point or line) located in the vicinity of loss-free structures. At this point, one cannot help wonder whether the whole question of existence and reality of guided waves is more a matter of definitions and the order of the magnitudes of the quantities involved, rather than distinct physical differences; this being the case, what is the right way of interpreting the mathematical expressions involved?*

The subdivision of a dynamic electromagnetic field into radiation field and guided waves is purely arbitrary and the only justification for this procedure lies in convenience of description, formulation of a clearer physical picture, etc. Although all electromagnetic waves are solutions to Maxwell's equations subject to the boundary conditions (and/or radiation condition at infinity) as well as launching conditions, the physical reality of any wave could be a matter of lengthy philosophical discussions.

The reason for this confusion is as follows: On the one hand, it can be argued that it is a necessary but not sufficient condition for the existence of a wave-type that it shall be independently a solution to Maxwell's equations subject to boundary and launching conditions; but it is only when, in addition, it can be shown that such boundary conditions and launching devices can in fact be physically realized, that the wave may be said to exist. On the other hand, it can be argued that the wave-type investigated need not on its own satisfy Maxwell's equations, but so long as it forms a part of a field, and the latter satisfies Maxwell's equations, boundary and launching conditions that the wave may be said to exist.

In short, the solution of the problem is characterized by complex integrals which contain pole singularities, for instance in problems of electromagnetic theory where the poles correspond to modes of a structure, and where one is interested in the effect of these modes on the total field of the excited structure. Examples of such modes could be Zenneck waves, lateral waves, leaky waves, or trapped slow surface waves. Although mathematical techniques for dealing with complex integrals are well developed, the interpretation of the results is not. The problem is associated with the presence of a pole that is accounted for in the saddle point evaluation of a complex integral as illustrated in [65] Figure 5. The various types of waves are described in the plot and their dependence of their position on the different branches of the various Riemann sheet. The resulting solution in the form of a contour integral is then examined in relation to the intrinsic properties of the media involved and the position of the antenna. The solution is obtained by splitting the integral into pole residues and branch-cut-integral; the latter is evaluated by developing it into a suitable asymptotic series [33-34]. The arbitrariness of the characterization is well stated by Schelkunoff: *the source of the difficulty involving the surface wave term in Sommerfeld's formulas ·was the double-valued nature of the square root terms in the reflection coefficient appearing in the integrand. In the mathematical formulation of the physical problem, it is essential that the square roots be assigned the values whose real parts are positive. Subsequent deformation of the contour of integration has to be conducted with great care and circumspection. No difficulty would have arisen if the deformation were made in a complex plane -with an impassable cut so that the square roots could take on only their principal values. As it happened, the deformation was made on a Riemann surface where it is quite easy for the reflection coefficient in the integrand to turn in to its reciprocal.*

Thus this integral-transform method is an important technique for treating radiation problems involving a source and plane homogeneous interfaces. The solution is in the form of an inverse transform-integral form, which is a spectral representation of the solution. The saddle-point (steepest-descent) technique is a useful approximate method for evaluating such integrals, when one is interested in the solution in regions of space reasonably far from the source. If a pole of the integrand is captured in the path deformation, a residue term is added to the saddle-point term. This residue term, a 'guided complex wave', has the functional form of an inhomogeneous plane wave propagating at some angle to the interface [66,67]. Tamir and Oliner [68] have published an excellent paper discussing such complex waves in detail. Their article is concerned with 2-dimensional fields above a plane homogeneous interface due to line-source excitation. The fields are approximated to by the first term of the saddle-point asymptotic expansion (space wave) plus residue terms (complex waves) present owing to poles having been captured in the path deformation. Tamir and Oliner [68] point out that additional terms are required for angles of observation for which the steepest descent path comes close to the poles, but that such angles are excluded from their discussion. They also point out that these additional terms serve to provide continuous transition between two regions dominated by fields of different spatial variation, and that the additional terms have no effect on features considered in their discussion. Their concepts has been pictorially depicted in [65] which is shown as Figure 5.

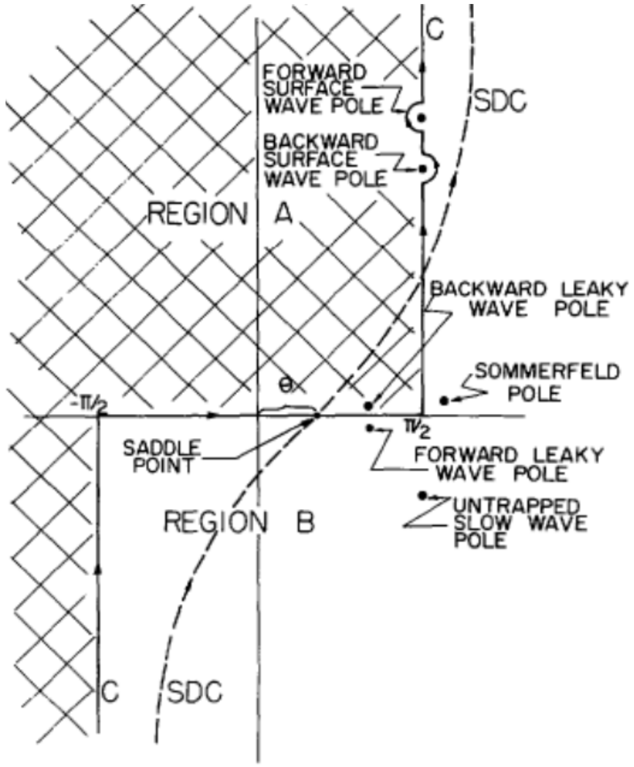


Figure 5. Complex plane of integration depicting the saddle point path of integration along with branch points and the poles including their nature of the contribution.

So the bottom line is to look at the additional terms required when the saddle point path is close to a pole, and use these terms to form a criterion for validity of the approximation of space wave plus complex waves. The objective is to show where in space above the interface the complex wave is a valid expansion and significant portion of the total field, and how the boundaries of this region of space vary with the parameters. A simple familiar example, that of a travelling-wave aperture distribution, is a good illustration. The modified saddle-point technique provides the approximation of space wave plus complex waves which comes directly from the ordinary saddle-point technique which has been described in Appendix B. (The adjective 'ordinary' is used here to distinguish this technique from the modified saddle-point technique which has been presented in Appendix C.) Under the ordinary saddle-point approximation, the effect of a pole is a residue term in its wedge of contribution, i.e. the complex wave [33]. Under the modified saddle-point approximation of Van der Waerden, the effect of a pole is a complementary- error-function term and modified coefficients of the saddle-point asymptotic expansion as illustrated in [33, 65-69].

The variety of waves which can be supported by a plane homogeneous interface includes surface waves of the forward and backward type, and several kinds of complex wave, the latter being characterized by wave numbers which are complex even though the media involved are not necessarily lossy as shown in Figure 5. All these waves are viewed as contributions due to poles in several alternative integral representations of a source-excited field, and places particular stress on the steepest-descent representation. Distinctions is made between proper (spectral, modal) and improper waves, and between lossy and lossless structures; complex waves along lossless structures are shown to appear always in degenerate pairs consisting of a forward and a backward wave as illustrated in Figure 5. There are discrete modes of propagation along guiding systems for both closed and open (radiating) structures. In lossless conventional waveguides the modes represent waves which either propagate without attenuation or decay exponentially without any phase change. Along open lossless structures, on the other hand, a continuous spectrum of modes is found, in addition to a discrete set, if any; the latter accounted for the various types of surface waves, while the continuous spectrum had a direct bearing on radiation. Both open and closed configurations can be characterized by propagation coefficients which are either purely real or purely imaginary, and this was demonstrated to apply to all structures which contain loss-less, isotropic and dispersionless media only again as presented in [63-69].

Waves associated with poles on the right branch are termed real waves and those associated with the wrong branch are the virtual waves [63-69]. In this way, real surface waves are slow waves and virtual surface waves are fast waves. But both are true guided waves. Waves associated with tubular metallic waveguides lose power laterally and can be termed “leaky waves.” These can be subdivided into evanescent and propagating modes, and provided the losses are not too large, the subdivision has practical significance. In this section we characterize the nature of the transmitted propagating wave by labelling it with a name to illustrate its behavior. In a cellular wireless communication system the radio wave propagation takes place through the Zenneck wave, which has appeared many times in different contexts in the various scientific literature. And in some places it may also appear as a lateral wave as we shall see. Next, the various types of waves are now defined. It is important to note that as Figure 5 illustrates the type of the various waves is directly determined as to what Riemann sheet they lie on and are therefore a function of the chosen location of the branch cuts. Hence their interpretations are quite arbitrary!

For a detailed explanation, consider a transverse magnetic (TM) plane wave that is incident at an angle θ with respect to the normal at the interface on the boundary between two semi-infinite, nonmagnetic, media separated by a planar boundary with the magnetic field parallel to the interface. The two media are air and a material medium having a permittivity ε. Part of the incident wave will be reflected and part of it will be transmitted through the interface. The TM reflection coefficient termed Γ_{TM} will be given by [68,19,34] (38) and is repeated here

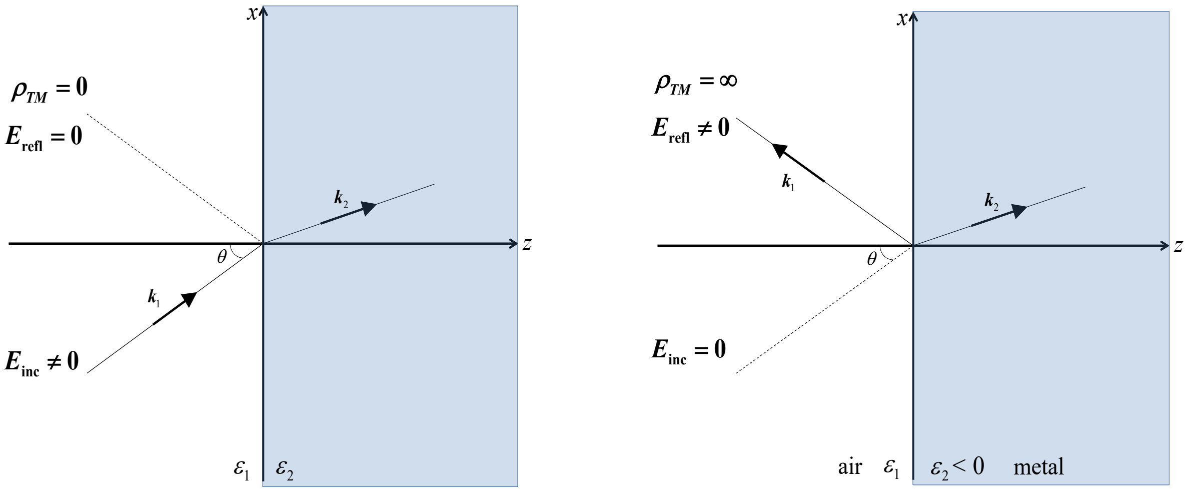
$$\Gamma_{TM} = \frac{\varepsilon \cos\theta - \sqrt{\varepsilon - \sin^2\theta}}{\varepsilon \cos\theta + \sqrt{\varepsilon - \sin^2\theta}} \tag{60}$$

At a grazing angle when $\theta \rightarrow 90°$ then $\Gamma_{TM} \rightarrow -1$. Also, there exists an incident angle at which the transmission is total and that is called the Brewster's angle. At the Brewster angle [34,68-70] defined by $\tan\theta_B = \sqrt{\varepsilon}$ and the reflection coefficient $\Gamma_{TM} = 0$ and there will be an incident field and a transmitted field but no reflected wave as shown in Figure 3.33a [70, p.270] and in Figure 6a. This is also the case for the Zenneck and Sommerfeld wave types. The Brewster angle is independent of the frequency of the incident wave and occurs at a *zero of the TM reflection coefficient.*

At a zero of the denominator of the reflection coefficient $\Gamma_{TM} = \infty$, there will be a reflected field and a transmitted field without any incident field as shown in Figure 3.33b [70, p.272] and in Figure 6b. This results in the generation of the surface wave. Therefore, neither a Zenneck wave nor a surface wave can be generated using an incident transverse Electromagnetic (TEM) wave but rather **only** through using TM waves which has a longitudinal component of the electric fields. In addition, a surface wave can only be generated quasi-particles or equivalently through an evanescent wave generated through a total internal reflection scenario. This is because in this case there is no incident wave and the evanescent wave is the cause of the response. This is because an evanescent wave behaves as a quasiparticle and can tunnel through the medium [70] exciting the electrons (this phenomenon of tunneling is called the Hartmann effect [71]). A true surface wave is generated by either the Otto [72] or the Kretschmann–Raether [73] methodology of using total internal reflection where evanescent waves are produced using a TM excitation under total internal reflection. The advantage of an evanescent wave is that it behaves as a quasiparticle and can tunnel through the metal (known as the Hartman effect [71], exciting the electrons to form a Surface Plasmon Polariton (SPP) as demonstrated in [72] and [73]. Hence, to generate a true surface wave no incident electromagnetic wave is necessary to generate the transmitted and the reflected fields as per Figure 6b. Also one needs a TM wave configuration and not a TEM one! For a surface wave to occur, alternately, the dielectric constants of the two medium should be of opposite sign and with the dielectric medium having a small loss, compared to the real part. A negative value for the dielectric constant with a small imaginary part occurs in metals generally in the petahertz (and not in terahertz) regions.

If the material medium in Figure 6a has a small dissipation then $\varepsilon = \varepsilon_r - \dfrac{j\sigma}{\omega\varepsilon_0}$ where ε is the complex permittivity of the medium. ε_r and ε_0 are the real permittivity of the medium and of vacuum, respectively. σ is the conductivity of the medium. ω is the frequency of the incident wave. For low loss, i.e., $\sigma / \varepsilon_r \omega \varepsilon_0$ □ 1 the effect of σ on the Brewster angle turns out to be of the second order [11,19,70]. If the incident waves are uniform plane waves, that is, their amplitude is constant in any plane perpendicular to the rays, there will be a small reflection [11,19,70]. In this case, the Brewster angle is the angle of minimum reflection. However, if the incident plane waves are inhomogeneous then this will be the condition for no reflection. It is known that the rays of inhomogeneous plane waves in a vacuum are straight lines along which the phase change is maximum and the amplitude remains constant. The planes perpendicular to the rays are equiphase planes [34,70]. Along these planes there exists one direction in which the amplitude does not vary, while in the perpendicular direction the amplitude change is maximum. The exponential rate of attenuation with the increase of distance is small but it is independent of the frequency. This is the case for the Zenneck wave. And indeed if one plots the variation of the reflected field in the air medium then it will not change as a function of frequency. However, for a surface wave as the frequency increases the variation of the fields will become more concentrated near the boundary [11,19,74]. One can demonstrate [19,74] that indeed this is the case for the cellular wireless communication as propagation takes place through the Zenneck/Sommerfeld wave type and not through a surface wave.

Next, we look at the various unique properties of each of these waves



a) Reflection Coefficient is zero: Brewster-Zenneck wave (b)) Reflection Coefficient is Infinity: Surface wave

Figure 6. Brewster-Zenneck (TM reflection coefficient $\rho_{TM} = 0$) and surface plasmon (TM reflection coefficient $\rho_{TM} = \infty$) cases for a two media problem.

4.1 Properties of a Zenneck Wave

Wait [2] states that the idea that a Zenneck type of wave could be employed to investigate the propagation of fields over earth was pointed out both by Zenneck himself and his colleague Hack [75]. In 1907 Zenneck [14,15] showed that a plane interface between two semi-infinite media such as the ground and air could support an electromagnetic wave which is exponentially attenuated in the direction of propagation along the surface and vertically upwards and downwards from the interface. Zenneck showed that Maxwell's equations did provide a solution for a wave or an inhomogeneous plane wave that was attached to the interface between air and the underlying medium and could propagate over great distances with a small amount of attenuation [15]. Zenneck did not show that an antenna could generate such a wave, but because this "surface wave" seemed to be a plausible explanation of the propagation of radio waves to great distances, it was accepted [11].

The main contribution of Zenneck was the development of a specific type of solution of Maxwell's equations in a three dimensional space [11,34,76,77]. This solution is an inhomogeneous type of plane wave and generally occurs at a zero of the TM reflection coefficient – the Brewster angle – whose field components can be derived as follows: In the three dimensional rectangular coordinates, consider that the plane $y = 0$ is the boundary between medium 2, free space (upper medium), and medium 1 (lower medium) which is of arbitrary parameters $(\varepsilon, \mu, \sigma)$ and the two media are homogeneous. Zenneck [14,15,19,34,70] showed that there exists a solution for Maxwell's equation in this two-layer problem. This solution represents a wave that has a progressive phase propagation in the *x*-direction, while at the same time decays exponentially in the positive and negative *y*-directions. This wave has to be a TM wave with respect to the *x-z* plane. The field components for such a wave in medium 1 are given by: [76-79]

$$\begin{aligned}
\mathrm{H}_{1z} &= A\ \exp(j\,\omega t - \gamma y + k_{1y}y) &&, y<0 \\
\mathrm{E}_{1x} &= A\left(\frac{k_{1y}}{\sigma + j\,\omega\,\varepsilon_1}\right)\exp(-k_x x + k_{1y}y) &&, y<0 \\
\mathrm{E}_{1y} &= A\left(\frac{k_x}{\sigma + j\,\omega\,\varepsilon_1}\right)\exp(-k_x x + k_{1y}y) &&, y<0
\end{aligned} \tag{61}$$

where $k_x = \alpha + j\beta$; A is a constant and $k_{1y} = \alpha_1 + j\beta_1$. Representing an attenuation α_1 and phase change β_1 for a wave travelling inwards from the surface. An operating frequency ω is assumed. The corresponding forms in medium 2 are given by: [76-79]

$$\begin{aligned}
\mathrm{H}_{2z} &= A\ \exp(j\,\omega t - k_x x - k_{2y}y) &&, y\ge 0 \\
\mathrm{E}_{2x} &= -A\left(\frac{k_{2y}}{j\,\omega\,\varepsilon_0}\right)\exp(j\,\omega t - k_x x + k_{2y}y) &&, y\ge 0 \\
\mathrm{E}_{2y} &= A\left(\frac{k_x}{j\,\omega\,\varepsilon_0}\right)\exp(j\,\omega t - k_x x + k_{2y}y) &&, y\ge 0
\end{aligned} \tag{62}$$

where $k_{2y} = \alpha_2 - j\beta_2$

Because the fields not only decays at the rate of α_2 with increasing distance from the surface but also suffers a progressive phase change β_2 for a wave travelling towards the surface. Here the power flow has two components, one representing the main stream along the interface and subject to the usual attenuation α with phase change β, while the other generally a minor one, is directed into the surface to supply the losses and so radiation therefore occurs. On both sides of the interface we have similar formula and thus within the surface $k_x^2 + k_{1y}^2 = j\omega\mu_0(\sigma_1 + j\omega\varepsilon_1)$ and $k_x^2 + k_{2y}^2 = -\,\omega^2\mu_0\,\varepsilon_0$ outside the surface in the air [77].

The power associated with a traveling wave is directed at an angle of inclination with the normal to the surface represented by the real part of the Brewster angle [11,34]. Thus the effect of the losses in the surface is to tilt the wavefront forward and to produce a decay of amplitude as the wave progresses over the surface. According to Zenneck, the real parts of $k_{1y}$and k_{2y} should be positive. Thus the wave given in (61)-(62) decays exponentially away from the boundary which lies at $y = 0$ plane, but is oscillatory in addition. The dispersion relation relating the propagation vector components is determined by the value of the frequency ω and the parameters of the medium $(\varepsilon, \mu, \sigma)$. The field of a Zenneck wave decays exponentially in amplitude and suffers a progressive advance in phase with increasing distance above the surface. Also, the wave is attenuated in the direction of propagation whilst subject to a progressive lag in phase along the interface [77].

Booker and Clemmow [80], in interpreting the Sommerfeld theory of propagation over a flat earth in the presence of finite losses, invoked the aid of a Zenneck wave. They discussed particularly the 2-dimensional case of a horizontal line source above earth. In doing this, they showed that a hypothetical Zenneck wave, when diffracted under a vertical plane screen, whose lower edge coincided with the image in the earth of the line source and extending to +∞ in the other direction as shown in Figure 2 [80, p.19], made equivalent provision for the effect of the earth losses. In Figure 2 [80, p.19], the field in the upper half space is generated by the sum of the fields produced by a line source located at the image point of the original source and the diffracted waves generated by a Zenneck wave incident on a screen without the presence of the earth as shown also in Figure 2 [80, p.19],. Booker and Clemmow [80] were very careful to point out that this was merely a convenient physical interpretation and that it did not mean that a Zenneck wave actually existed above the earth's surface, but it appears nevertheless that there has been some misunderstanding which can, no doubt, be attributed to the use of the term surface wave for two different things.

As an example, consider the propagation of a wave radiated by a half wave dipole located at a height of 10 m over an urban ground as illustrated in [19,74] where the radio wave propagation is through a Zenneck wave. This is close to a TM wave incident at the air-Earth interface. We now plot the variation of the magnitude of the **reflected field only** as a function of z for a fixed distance of ρ = 100 m from the transmitter. The vertical variation of the reflected field as a function of the height of upto 100 m from the ground is seen in the plot of Figure 3.34 of [19]. It is seen that even though the frequency changes from 453 MHz, to 906 MHz and then to 1350 MHz, the magnitude of the decay of the reflected fields in the vertical direction to the planar air-earth interface remain practically the same. The fields tend to decay monotonically resembling an evanescent wave. As there is no appreciable variation of the fields even though the frequency changed by a factor of three, indicates that the wave in this case relates to a Zenneck wave and not to a surface wave as correctly suggested by Schelkunoff [11]. This is applied to the waves propagating over a two-layer medium like radio wave propagation in an urban environment. Also, in the plot of Figure 3.34 of [19,74] in the absolute value of the reflected fields, there is an indication of a dip in the strength of the reflected fields at a height of 40 m from the ground. Considering urban ground which has a $\varepsilon_r = 4$, will yield a Brewster's angle given by $\tan^{-1}(\sqrt{4}) = 63.4°$. From the geometry of the problem, it is seen that the angle subtended by the antenna with respect to the ground will be given by $\tan^{-1}(10/20) = \tan^{-1}(40/80) \approx 26.6°$. Therefore, the angle with respect to the vertical is $90° - 26.6° = 63.4°$, indicating that the dip in the field strength is occurring exactly at the Brewster angle of 63.4° even though the permittivity of the ground is complex. This indicates as Schelkunoff [11] had predicted that the effect of small conductivity on the Brewster's angle is of the second order and the location of the Brewster's angle does not change with frequency even though it varies over a factor of 3 (from 453 MHz to 1350 MHz).

The Zenneck wave is an evanescent fast wave. Next we look at the concept of a true surface wave.

4.2 WHAT IS A SURFACE WAVE?

As explained by Barlow [76] and Barlow and Cullen [77], and further illustrated in [78,79], a *surface wave* is one that propagates without radiating energy along an interface between two different media. If both media have finite losses, the energy directed along the interface will be required to supply the losses in the media. This does not invalidate the description of the surface wave if radiation is construed to mean that energy is absorbed from the wave independent of the media supporting it. The surface wave phenomenon arises primarily from its unique non-radiating characteristic which enables high frequency energy to be transferred intact from one point to another, except in so far as demands are made upon that energy to compensate for the losses in the two media. This definition agrees with the concept of Rayleigh as we discussed in the previous section. However, the next statement of Barlow and Cullen [77] which states that the three distinctive forms of the surface wave, namely

1) the Zenneck or inhomogeneous plane wave supported by a flat surface,
2) the radial cylindrical wave also supported by a flat surface, and

3) the Goubau [60,81] or axial cylindrical wave associated with a transversely cylindrical surface, represent basically one and the same phenomenon, from their field distribution, and this can cause some anxiety as we will soon see.

In addition, Barlow et al. [76-79] pointed out that Sommerfeld's theory for ground wave propagation over a flat earth also introduced a so called *surface wave*. Sommerfeld divided the ground wave into two components, which he called respectively a "space wave" and a "surface wave". The surface wave component is represented by one of the terms in the analysis of the total field, and its particular feature is that it needs to predominate near the earth's surface [11]. Both parts are required to satisfy Maxwell's equations, and at long ranges according to Sommerfeld the surface wave part varies inversely as the square of the distance, which is identical to the Norton surface wave. In similar circumstances, a true surface wave radiated from a vertical line source over a horizontal surface would be expected to decay exponentially with range, owing to losses and at the same time fall in amplitude inversely as the square root of the range, due to the expanding wave-front. In Sommerfeld's original 1909 paper a surface wave of this type appeared in the expression of the total field, which was later shown not to exist [81].

In addition, Wait [82] also provides a clarification between the terms surface waves and ground waves [34,82,83], which are the fields observed at the interface. According to Wait: "A surface wave is one that propagates along an interface between two different media without radiation; such radiation being construed to mean energy converted from the surface wave field to some other form", according to the definition of Barlow and Cullen [77] and Cullen [79]. The *ground wave* is characterized as per the IEEE definition [35,84]: "A radio wave that is propagated over the earth and is ordinarily affected by the presence of the ground and troposphere. The ground wave includes all components of a radio wave over the earth except ionospheric and tropospheric waves".

The situation is particularly confused since in the IEEE Test procedures, the surface wave component of the ground wave is completely different from the definition used by Barlow and Cullen [77]. Wait [83] tries to clarify this state of confusion by adopting a model which would encompass all forms of waves which can propagate over an interface. This was carried out by expressing the total field from a vertical electric dipole at a height *h* radiating over an air-earth interface. The vertical electric field at a vertical height *z* and a horizontal distance ρ could be written in the following form consisting of three terms:

$$E = E_a + E_b + E_s$$

where E_a is the field which would be computed on the basis of geometrical optics, whereas E_b and E_s are the *surface waves* which are entirely different in character from one another. If the phase angle of the surface impedance is less than 45° [83], it was shown that E_s was not present. Then E_b could be identified as a correction to the geometrical optics field E_a. Asymptotically, E_b varies as ρ^{-2} whereas E_a varies as ρ^{-1}. Thus, in many high frequency applications, E_a is dominant. However, at shorter distances and/or lower frequencies when the asymptotic form for *E* is not valid, it turns out that E_b may be very important and that it should be called a Norton surface wave [35, 43-45,83], rather than just referring to it as a surface wave.

When the phase angle of the surface impedance becomes greater than 45°, the contribution E_s is finite and in many cases, it dominates E_a and E_b. For example, it may be shown that [83] E_s varies as $\rho^{-1/2}$ for a purely inductive boundary. The contribution E_s, which is not present for a homogenous conducting half space, is really a trapped surface wave since the energy is confined to regions near the interface. It is suggested that this component of the total field be described as the *Barlow surface wave*. It is of interest to note that, at least formally, E_s has the form of a Zenneck wave for the case of a homogenous earth. As discussed in [85] it is not excited by a physically realizable source.

From these various interpretations of the term surface waves, it is very easy to get confused. At this point, we take recourse to Schelkunoff [11] as he clarifies the situation from a mathematical perspective and presents the definition of what a *surface wave* is in the true classical sense of Rayleigh. Schelkunoff [11] points out that these same words *surface wave* convey different meanings to different individuals. According to Schelkunoff, *the preliminary list of different surface waves compiled by Dr. James R. Wait, the then chairman of a working group of URSI, mentions 11 types of surface waves in the light of propagation of plane waves in two semi-infinite, nonmagnetic, nondissipative media, separated by a plane boundary. These 11 types of surface waves consist of*

1. *Zenneck Surface Wave* (interface of two half-space having different dielectric constants),
2. *Sommerfeld Surface Wave* (dipole over a conducting half-space),
3. *Norton Surface Wave* (dipole over a conducting half-space),
4. *Sommerfeld Axial Surface Wave* (imperfectly conducting cylindrical wire),
5. *Goubau Axial Surface Wave* (dielectric-coated wire),
6. *Plane Trapped Surface Wave* (dielectric-coated plane conductor, corrugated surface, or other inductive boundaries),
7. *Cylindrical Trapped Surface Wave* (same as above in cylindrical form),
8. *Plane Quasi-Trapped Surface Wave* (stratified conductor when the surface impedance has both a resistive and inductive component),
9. *Cylindrical Quasi-Trapped Surface Wave* (same as above in cylindrical form),
10. *Azimuthal Surface Wave* (on dielectric-coated and corrugated cylinders and spheres for propagation in the azimuthal direction),
11. *Composite Axial-Azimuthal Surface* Wave (same as above when propagation has a component in both the axial and azimuthal directions).

As a group these wave types have no important physical properties in common. Calling these wave types by the same name, even with qualifying adjectives, encourages one to assume that the most significant physical properties of one wave type are shared by other wave types and can cause serious misunderstandings. As Schelkunoff explains [11] when a wave is incident at an air-dielectric boundary then there can be partial transmission and partial reflection, in the form of radiation, from the boundary. However, if the wave contains a component whose magnetic field is parallel to the boundary and the wave is incident at a Brewster's angle then the wave will be totally transmitted to the dielectric, and if there is a reflected wave it will be horizontally polarized, as the Brewster's angle apply only to vertically polarized waves. Thus, Brewster's angle is associated with the zeros of the reflection coefficient (zeros of the numerator of the term placed inside brackets in (19)). If the medium is slightly lossy then again there will be minimal reflection from the boundary. In this case, the wave is not tied to the boundary. The Brewster's angle is independent of the frequency and so is the exponential rate of attenuation of this wave with the increasing distance from the interface, which is small. The phase constant in vacuum increases with the decrease in the wavelength for this wave [11]. If the dielectric medium has a finite conductivity then there may be an attenuation at right angles to the direction of propagation. Also, the wave is a fast wave and essentially passes through the dielectric medium without seeing it. The first four of the eleven wave types, listed above, belong to this class and therefore are not true surface waves in the classical sense of Lord Rayleigh [11].

However, when a wave is incident at a boundary from a denser to a rarer medium then there will be total internal reflection if the incident wave is incident at an angle equal to or greater than the critical angle of the medium. The wave may even get trapped in the denser dielectric medium and may never come out except providing evanescent field to the rarer medium. The wave in this case is a slow wave and the evanescent fields in the rarer medium will get concentrated near the interface as the frequency increases [11,19,70]. The attenuation in the vertical direction away from the interface in the rarer medium will also be frequency dependent and as the frequency increases the wave will be confined to the interface. Such a situation occurs for the poles of the reflection coefficient (zeros of the denominator of the term placed inside the brackets in (19)). In this case, even when the incident field goes to zero the amplitude of the surface waves does not decrease. Typically, these waves do not have a decay along the direction of propagation in contrast to a Zenneck wave, but if the dielectric medium is lossy there may be an anomalous velocity of propagation towards the interface [11]. Physically these types of trapped waves may be stirred up by the incident field. According to Schelkunoff waves type 5-11 listed above, satisfy the characteristics of a *true surface wave* in the classical sense introduced by Lord Rayleigh. Thus, the Zenneck wave is not a true surface wave in the classical sense [11].

Surface plasmons are excited in metal foils of special thickness by quasi particles and not by a transverse electromagnetic wave since these surface waves are TM waves with a longitudinal field component. Surface plasmons should not be related to Zenneck waves and the exciting conditions for the Surface Plasmon Polaritons (SPP) should be strictly defined. Unfortunately, a clear statement that SPP wave is indeed a TM surface wave with a longitudinal field component and not a TM Zenneck wave is missing in most recent works on terahertz surface plasmons. SPPs are true surface waves with a longitudinal field component and are generated when the two medium has permittivity of opposing signs. However, this is not sufficient for the existence of the pole as many metals has a negative permittivity at terahertz frequencies but the conductivity is still too large. At petahertz frequencies for metals when the conductive losses become smaller than the permittivity then only will the surface wave pole of the TM reflection coefficient will manifest itself. Surface waves are thus in no way related to a Zenneck wave which is generated due to a Brewster zero – the zero of the TM reflection coefficient

4.3 WHAT IS A LEAKY WAVE?

Leaky waves are solutions of the source-free Maxwell equations and are not proper modes because of their singular behavior at infinity in the transverse, or cross section, plane. Despite the decidedly unphysical nature of this aspect of the leaky-wave solutions, however, they are valid field representations in certain restricted regions within which they remain finite. It is the purpose of this section to expand on these remarks. [63,64].

The radiation from a continuous longitudinal slot in a uniform lossless waveguide can frequently be characterized by a traveling wave with a complex propagation constant. This traveling wave propagates along the waveguide with a velocity greater than that of light and is attenuated as it travels, thus indicating a continuous leakage of energy. Waves possessing this characteristic behavior have been designated as leaky waves. These leaky waves are not characteristic modes of the open waveguide region. The nonmodal character of such waves is evident from the fact that they increase without limit in the transverse direction. This improper behavior of the leaky waves and the distinction between them and characteristic modes are discussed next. Despite this unphysical behavior, however, these nonmodal waves may nevertheless be employed for the representation of field solutions in suitably restricted portions of open waveguide regions. Moreover, since the leaky waves are solutions of the source-free field equations, their propagation constants can be obtained rigorously from a transverse resonance calculation which is formally identical to that for discrete characteristic modes [63-69].

A more careful enumeration of the characteristics of the leaky waves puts them in sharp contrast to surface waves, with which there has been considerable confusion [63-69]. A leaky wave is not a proper modal solution, whereas the surface wave is. The surface wave propagates unattenuated parallel to the guiding surface and has pure attenuation transverse to the surface. The leaky wave, in contrast, has complex propagation constants in both longitudinal and transverse directions, attenuating in the forward phase direction along the surface and growing transverse to the surface. The surface wave is a slow wave, whereas the leaky wave has a longitudinal phase velocity greater than that of light. Finally, the surface wave cannot radiate, whereas the leaky wave is not bound to the surface and may therefore contribute to the radiation field of the structure. Although, in general, leaky waves exist on structures that are quite different from those supporting surface waves, there are some structures that support both types of waves. The existence of leaky waves [66] on a surface interface structure has been postulated by Zucker [86]. It was subsequently found that, in the absence of sources at finite distances, leaky waves would exist in a physical sense only if the guiding structure were modulated periodically. [87-89]

The nonmodal nature of leaky waves can be exhibited only by a careful examination of the modal spectrum of the guiding structure. The modal spectrum of an open waveguide is usually purely continuous although under certain conditions there may also exist a discrete spectrum. It can be shown that the complete modal spectrum is given by an integration over all the singularities of an appropriate one-dimensional Green's function, called the characteristic Green's function. In closed waveguides, only pole singularities are present and the resulting modal spectrum is purely discrete. In open regions, branch point singularities also occur, and the integration for the modal spectrum must be performed on the proper branch, the one corresponding to decaying waves at infinity. The residues of the poles, if any, on the proper branch then correspond to true discrete modes, while the branch cut integration(s) yields the continuous spectrum. Also, pole singularities are often present on the wrong branch, and in the past their presence, if noted, have been ignored since these poles did not contribute to the spectrum and were unphysical in that they corresponded to growing waves at infinity. However, Marcuvitz [63] recognized that these poles are in fact leaky waves, that under certain conditions some of them may contribute to the radiated field, and that they are valid only in certain suitably restricted regions. In the presence of a source, it has been found that leaky waves also exist on unmodulated structures. Marcuvitz [61] has predicted the existence of nonmodal waves due to a source exciting a uniform guiding surface, which have characteristics identical in form to the waves referred to above. Marcuvitz [61] points out that such waves are, in reality, part of the continuous spatial spectrum of the source and are distinct from the discrete (modal) spectrum of surface waves. Furthermore, these waves are exponentially attenuated along any radius from the source within their domain of existence and therefore comprise a part of the near field, so to speak, of the source.

Interest in such waves should not be limited to the theoretical domain since the design of excitation structures for surface-wave transmission lines, antennas and cavities should take these waves into account if there are obstructions to or changes in the surface-wave structure [81].

4.4 WHAT IS A LATERAL WAVE?

An alternative formulation is discussed wherein the interface effects are accounted for one at a time and the resulting diffraction field is then shown to involve lateral waves (branch-cut waves). The two representations are compared and their respective utility is illustrated by examples. When the source and observation points are located exterior to a large dielectric gap, diffraction effects due to an accumulation of leaky waves are found to be equivalent to a single lateral wave. For source and observation points inside a lossy dielectric slab, the pole-wave formulation provides a somewhat more convenient but physically less transparent result than the one comprising lateral wave. [69].

If the source and observation points are located outside the slab and are separated by a distance of many wavelengths, then it is well known that the field is comprised of a direct and reflected wave which together constitute the geometric-optical field, and also has a diffracted contribution. The diffracted field has customarily been given in terms of surface waves and (or) leakv waves which may be supported by the structure and may enter into the solution in certain restricted spatial regions. This formulation results from a modal analysis of the problem wherein a composite reflection coefficient is employed to account for the effect of the slab on a typical plane-wave modal constituent; the source field is then synthesized by modal superposition as stated in [90]. Also as illustrated by Tamir [91]. Since the lateral waves are generally closely associated with the phenomenon of total reflection, their influence is most pronounced when the source and observation points are both situated in the denser dielectric. A different representation will thereafter be derived wherein the discrete pole-wave spectrum is replaced by a finite number of lateral (branch-cut) waves. The study of these alternative formulations of diffraction effects forms the basis of [69,90].

Finally, examining (52) and (56) it is seen that the Green's function for the Sommerfeld problem has a variation of distance as $R^{-1.5}$ which points to the direction of the presence of a lateral wave as such a variation with distance is the hall mark of a lateral wave. The continuous leakage of energy at the interface is reflected in an amplitude variation of $R^{-1.5}$ rather than $R^{-0.5}$ indicating that on portions of the path wireless propagation can be seen by a lateral wave [69]. However, in wireless propagation the source is in the air and typical lateral waves often also referred to as branch cut waves occur when radio waves traverses form a denser to a rarer medium [69,90,91]. To continue, van der Pol pointed out that the semi-infinite Sommerfeld integrals can be modified analytically to an integral of the following form [92,93,19]. As shown in [92,93,19] Vander Pol analytically modified the Sommerfeld integrals to reflect that the fields in the upper medium can be produced due to an image of the original dipole in the ground and as stated: *The integration in this new form extends over the part of the entire spatial volume occupied by the second medium below the geometrical image of the source. The fields in the first medium apart from the direct radiation from the original source can be described now due to secondary waves originating in the integration space below the image as mentioned. The amplitude of these secondary waves being determined by the amplitude of a primary wave which can be considered to spread from the geometrical image of the source with a propagation constant and absorption belonging to the second medium*. This indicates that the effect of the field generated by the image of an dipole radiating over an imperfect ground can be interpreted as energy coming from the denser medium to the rarer medium and in some regions can be identified as a lateral wave [90].

Next a partial list of relevant references are provided. The list is by no means complete. It presents related references where additional materials can be obtained. This followed by three Appendices.

VII References

[1] J. C. Maxwell and W. D. Niven, A Treatise on Electricity and Magnetism, 3rd Edition, 1954.

[2] J. R. Wait, "The Ancient and Modern History of EM Ground Wave Propagation," *IEEE Antennas and Propagation Magazine*, Vol. 40, No. 5, October 1998, pp. 7–24.

[3] R. E. Collin, "Hertzian Dipole Radiating over a Lossy Earth or Sea: Some Early and Late 20th-century Controversies," *IEEE Antennas and Propagation Magazine*, Vol. 46, No. 2, April 2004, pp. 64–79.

[4] H. Hertz, *Electric Waves*, Macmillan & Co., London, UK, 1893.

[5] R. Appleyard, *Pioneers of Electrical Communication*, Macmillan Company, New York, NY, USA, 1930.

[6] J. G. Growther, *Six Great Inventors*, Hamish Hamilton, London, UK, 1961.

[7] N. Tesla, *Art of Transmitting Electrical Energy through Natural Mediums*, US Patent 787412, April 18, 1905.

[8] http://en.wikipedia.org/wiki/Alexander_Stepanovich_Popov.

[9] T. K. Sarkar and D. L. Sengupta, "An Appreciation of J. C. Bose's Pioneering Work in Millimeter Waves," *IEEE Antennas and Propagation Magazine*, Vol. 39, No. 5, Oct. 1997, pp. 52–62.

[10] G. Marconi, "Wireless Telegraphy," *Jour. IEE* (UK), Vol. 28, 1899, pp. 273–315.

[11] S. A. Schelkunoff, "Anatomy of '"Surface Waves"'," *IRE Trans. on Antennas and Propagation*, Vol. 7, No. 5, Dec. 1959, pp. 133–139.

[12] E. Cohn, *Das Elekromagnetische Field-Vorlesungen über die Maxwell'sche Theorie*, Leipzig, Germany, 1900.

[13] K. Uller, *Ph.D. Dissertation*, Univ. of Rostock, Germany, 1903.

[14] J. Zenneck, "Propagation of Plane Electromagnetic Waves along a Plane Conducting Surface and Its Bearing on the Theory of Transmission in Wireless Telegraphy," *Ann. Phys*., Vol. 23, Sept. 1907, pp. 846–866.

[15] J. Zenneck, *Wireless Telegraphy*, McGraw-Hill Book Co., New York, NY, USA, 1915.

[16] A. N. Sommerfeld, "Propagation of Waves in Wireless Telegraphy," *Ann. Phys*., Vol. 28, Mar. 1909, pp. 665-736.

[17] A. N. Sommerfeld, "Propagation of Waves in Wireless Telegraphy," *Ann. Phys*., Vol. 81, Dec. 1926, pp. 1135 - 1153.

[18] B. Rolf, "Graphs to Prof. Sommerfeld's Attenuation Formula for Radio Waves", *Proc. IRE*, Vol. 18, 1930, pp. 391-402.

[19] T. K. Sarkar, M. Salazar and A. Abdallah, *The Physics and Mathematics of Radio Wave Propagation in Cellular Wireless Communication*, IEEE Press and John Wiley & Sons, May 2018.

[20] J. R. Wait, *Electromagnetic Waves in Stratified Medium*, IEEE Press, 1996.

[21] J. A. Stratton, *Electromagnetic Theory*, McGraw-Hill Book Company, New York, 1941, p. 573.

[22] A. N. Sommerfeld, *Partial Differential Equations in Physics*, Academic Press, New York, 1949.

[23] L. Brekhovskikh, *Waves in Layered Media*, Academic Press, New York, 1973.

[24] A. Ishimaru, *Electromagnetic Wave Propagation, Radiation, and Scattering*, Englewood Cliffs, New Jersey, 1991

[25] A. Baños, Jr., *Dipole Radiation in the Presence of a Conducting Half-Space*, Pergamon Press, Oxford, England, 1966, pp. 151-158.

[26] G. Tyras, *Radiation and Propagation of Electromagnetic Waves*, Academic Press, New York, 1969.

[27] E. K. Miller, A. J . Poggio, G. L. Burke and E. S. Selden, "Analysis of Wire Antennas in the Presence of a Conducting Half Space: Part I. The Vertical Antenna in Free Space," *Canadian Journal of Physics*, Vol. 50, 1972, pp. 879-888.

[28] J. A. Kong, Electromagnetic Wave theory, EMW Publishing, Cambridge, MA, 1986, 2005, 2008.

[29] T. K. Sarkar, “Analysis of Arbitrarily Oriented Thin Wire Antennas over a Plane Imperfect Ground,” *AEU*, Band 31, Heft 11, 1977, pp. 449-457.

[30] G. K. Karawas, *Theoretical and Numerical Investigation of Dipole Radiation over a Flat Earth*, Ph. D. Dissertation, Case Western Reserve University, Cleveland, OH, 1985.

[31] T. K. Sarkar, W. Dyab, M. N. Abdallah, M. Salazar-Palma, M. V. S. N. Prasad, S. Barbin, and S. W. Ting, “Physics of Propagation in a Cellular Wireless Communication Environment,” *Radio Science Bulletin*, No. 343,

[32] P. C. Clemmow, *The Plane Wave Spectrum Representation of Electromagnetic Fields*, Pergamon Press, New York 1966.

[33] L. B. Felsen and N. Marcuvitz, *Radiation and Scattering of Waves*, Prentice Hall, New Jersey, 1973.

[34] R. E. Collin, *Field Theory of Guided Waves*, McGraw-Hill, 1960.

[35] “IEEE Standard Definitions of Terms for Radio Wave Propagation”, *IEEE Std* 211-1997 Publication Year: 1998

[36] Y. Okumura, E. Ohmori, T. Kawano, and K. Fukuda, “Field Strength and Its Variability in VHF and UHF Land Mobile Service,” *Rev. Elect. Commun. Lab.*, Vol. 16, No. 9–10, 1968, pp. 825–873.

[37] A. E. Kennelly, “On the Elevation of the Electrically Conducting Strata of the Earth’s Atmosphere,” *Elec. World and Engr*., Vol. 39, March 1902, p.473.

[38] O. Heaviside, *Encyclopedia Brittanica*, Tenth Edition, 1902.

[39] G. Brett and M. A. Tuve, “A Test of the Existence of the Conducting Layer,” *Phys. Rev*., Vol. 28, Sept. 1926, pp. 554–575.

[40] H. Weyl, “Propagation of Electromagnetic Waves over a Plane Conductor,” *Ann. Phys*., Vol. 60, 1919, pp. 481–500.

[41] B. van der Pol and K. F. Niessen, “The Propagation of Electromagnetic Waves over a Plane Earth,” *Ann. Phys*., Vol. 6, 1930, pp. 273–294.

[42] B. van der Pol, “Ueber die Ausbreitung Electormagnetischer Wellen,” *Jahrbuck der Drahtlosen Telepraphie Zeitschrift fuer Hochfrequenztechnik*, Vol. 37, 1931, pp. 152–156.

[43] K. A. Norton, “The Propagation of Radio Waves over the Surface of the Earth,” *Proceedings of the IRE*, Vol. 24, 1936, pp. 1367–1387.

[44] K. A. Norton, “The Physical Reality of Space and Surface Waves in the Radiation Field of Radio Antennas,” *Proceedings of the IRE*, Vol. 25, 1937, pp. 1203-1236.

[45] K. A. Norton, “The Physical Reality of Space and Surface Waves in the Radiation Field of Radio Antennas,” *Proceedings of the IRE*, Vol. 25, 1937, pp. 1192–1202.

[46] K. A. Norton, “Propagation of Radio Waves over a Plane Earth,” *Nature*, Vol. 135, 1935, pp. 954–955.

[47] K. F. Niessen, “Zur Entscheidung zwischen den Bieden Sommerfeldschen Formeln fuer die Fortpflanzeng von Drahtlosen Wellen,” *Ann. Phys.*, Vol. 29, 1937, pp. 585–596.

[48] C. R. Burrows, “The Surface Wave in Radio Propagation over Plane Earth,” *Proc. IRE*, Vol. 25, May 1937, pp. 219–229.

[49] C. R. Burrows, “Radio Propagation Over Plane Earth – Field Strength Curves”, *Bell Systems Technical Journal*, Vol. 16, 1937, pp. 45–75.

[50] K. L. Corum, M. W. Miller, and J. F. Corum, “Surface Waves and the Crucial Propagation Experiment”, Texzon Technology, Wireless Power conference 2016.

[51] W. H. Wise, “The Physical Reality of the Zenneck Wave,” *Bell Systems Technical Journal*, Vol. 16, 1937, pp. 35–44.

[52] S. O. Rice, “Series for the Wave Function of a Radiating Dipole on the Earth’s Surface,” *Bell Systems Technical Journal*, Vol. 16, 1937, pp. 101–109.

[53] A. N. Sommerfeld, Letter to K. A. Norton from the Zirmerhof, Tyrol, Sept. 5, 1937.

[54] K. A. Norton, Letters to Prof. Sommerfeld, both dated Sept. 20, 1937.

[55] K. A. Norton, “The Sommerfeld Error in Sign,” Annex to Lecture X.2, *EM Propagation Course*, Boulder, Colorado, USA, July 3, 1968.

[56] F. Noether, “Spreading of Electric Waves along the Earth,” in E. Rothe, F. Ollendorff, and K. Polhausen, *Theory of Functions as Applied to Engineering Problems*, MIT Technology Press, Cambridge, MA, USA, Part II, Ch. E, pp. 167-185.

[57] H. Ott, “Reflecxion und Brechung von Kugelwellen; Effekte 2. Ordung,” *Ann. Physik.*, Vol. 41, 1942, pp. 443–466.

[58] P. S. Epstein, “Radio Wave Propagation and Electromagnetic Surface Waves,” *Physics*, Vol. 33, 1947, pp. 195–199.

[59] C. J. Bouwkamp, “On Sommerfeld’s Surface Wave,” *Physical Review*, Vol. 80, No. 2, 1950, p. 294.

[60] T. Kahan and G. Eckart, “On the Electromagnetic Surface Wave of Sommerfeld”, *Physical Review*, Vol. 76, No. 3, 1949, pp. 406-410.

[61] T. Kahan and G. Eckart, “On the Existence of a Surface Wave in Dipole Radiation over a Plane Earth”, *Proc. Of IRE*, Vol. 38, No. 7, 1950, pp. 807-812.

[62] G. Goubau, “*Waves on interfaces*”, IRE Transactions on Antennas and Propagation, Vol. 7, No. 5, 1959, pp. 140-146.

[63] N. Marcuvitz, “*On field representations in terms of leaky modes or Eigenmodes”,* IRE Transactions on Antennas and Propagation, 1956, Vol. 4, Issue: 3, pp. 192 – 194.

[64] A. Karbowiak, “*Radiation and guided waves”,* IRE Transactions on Antennas and Propagation, 1959, Vol. 7, Issue: 5, pp. 191 – 200.

[65] G.D. Bernard and A. Ishimaru, “*On complex waves”, Proceedings of the Institution of Electrical Engineers*, 1967 , Vol. 114 , Issue: 1, pp. 43 – 49.

[66] L. Goldstone and A. Oliner, “Leaky Wave antennas I: Rectangular waveguides*”,* IRE Transactions on Antennas and Propagation, 1959 , Vol. 7 , Issue: 4, pp. 307 – 319’

[67] E.S. Cassedy and M. Cohn, “*On the Existence of Leaky Waves Due to a Line Source Above a Grounded Dielectric Slab*”, IRE Transactions on Microwave Theory and Techniques, 1961 , Vol. 9 , Issue: 3, pp. 243 – 247.

[68] T. Tamir and A. A. Oliner, “*Guided Complex Waves: Pt. 1*,” Proc. IEE (Brit.), Vol. 110, No. 2, 1963, p. 310-24.

[69] T. Tamir and L. Felsen, “*On lateral waves in slab configurations and their relation to other wave types*”, IEEE Transactions on Antennas and Propagation, 1965 , Vol. 13 , Issue: 3, pp. 410 – 422.

[70] S. J. Orfanidis, *Electromagnetic Waves and Antennas,* Rutgers University, Piscataway, N.J., 2012. http://www.ece.rutgers.edu/~orfanidi/ewa/.

[71] T. E. Hartman, “Tunneling by a wave packet”, *Journal of Applied Physics*, Volume 33, Issue 12, pp. 3427-3433, 1962.

[72] A. Otto, “Excitation of nonradioactive surface plasma waves in silver by the method of frustrated total reflection,” *Z. Physik*, vol. 216, pp. 398–410, 1968.

[73] H. Raether, *Surface Plasmons on Smooth and Rough Surfaces and on Gratings*, Berlin, Germany: Springer-Verlag, 1988.

[74] T. K. Sarkar, M. N. Abdallah, M. Salazar Palma and W. M. Dyab, “Surface Plasmons/Polaritons, Surface Waves and Zenneck Waves: Clarification of the Terms and a Description of the Concepts and their Evolution”, *IEEE Antennas and Propagation Magazine*, Vol. 59, No. 3, June 2017, pp. 77-99.

[75] F. Hack, “Propagation of EM Waves over a Plane Conductor,” *Ann. Phys*., Vol. 27, Sept. 1908, pp. 27–37.

[76] H. M. Barlow, “Surface waves”, *Proceedings of the IRE*, 1958 , Vol. 46 , Issue: 7, pp. 1413 - 1417

[77] H. M. Barlow and A. L. Cullen, “Surface Waves,” *Proc. of IEE*, Vol. 100, Part III, 1953, pp. 329–341.

[78] H. M. Barlow and j. Brown, *Radio Surface Waves*, Clarendon Press, Oxford, UK, 1962.

[79] A. L. Cullen, “Excitation of Plane Surface Waves,” *Proc. of IEE*, Vol. 101, 1954, pp. 225–234.

[80] H. G. Booker and P. C. Clemmow, “A Relation between the Sommerfeld Theory of Radio Propagation over a Flat Earth and the Theory of Diffraction at a Straight Edge”, *Proc. of IEE*, Vol. 97, Part III, No. 45, Jan 1950, pp. 18–27.

[81] G. Goubau, “Surface waves and their application to transmission lines,” *Journal of Applied Physics*, Vol. 21, 1950, p.1119.

[82] J. R. Wait, “Electromagnetic Surface Waves,” in *Advances in Radio Research*, Vol. 1, Academic Press, New York, NY, USA, 1964, pp. 157–217.

[83] J. R. Wait, “A Note on Surface Waves and Ground Waves,” *IEEE Trans. Antennas and Propagation*, Vol. AP-13, Nov. 1965, pp. 996–997.

[84] “IEEE Test Procedures for Antennas,” *IEEE Trans. on Antennas and Propagation*, Vol. AP-13, No. 5, May 1965, pp. 437–466.

[85] D. A. Hill and J. R. Wait, “Excitation of the Zenneck Surface Wave by Vertical Apertures,” *Radio Sci*., vol. 13, 1978, pp. 969-977.

[86] F. J. Zucker, "The guiding and radiation of surface waves," *Proc. Symp. on Modern Advances in J of Microwave Techniques,* Polytechnic Inst. of Brooklyn, Brooklyn, N. Y., November 8-10, 1954, vol. 4, pp. 403-435.

[87] A. S. Thomas and F. J. Zucker, "Radiation from modulated surface wave structures-I," 1957 IRE National Convention Record, pt. 1, pp. 153-160.

[88] R. L. Pease, "Radiation from modulated surface wave ·structures," 1957 IRE National Convention Record, pt. 1, pp. 161-165.

[89] A. A. Oliner and A. Hessel, "Guided waves on sinusoidally modulated reactance surfaces," IRE Trans. On Antennas And Propagation, vol. AP-7, pp. S201-S208; December, 1959.

[90] R. W. P. King, “Electromagnetic Field of a Vertical Electric Dipole over an Imperfectly Conducting Half-Space,” *Radio Science*, Vol. 25, March-April 1990, pp. 149-160.

[91] T. Tamir, “Experimental verification of a lateral wave above a lossy interface”, *Electronics Letters*, 1970, Vol. 6 , Issue: 12, pp. 357 – 358.

[92] B. Van der Pol, “Theory of the Reflection of Light from a Point Source by a Finitely Conducting Flat Mirror with Application to Radiotelegraphy,” *Physics*, Vol. 2, Aug 1935, pp. 843–853.

[93] T. K. Sarkar , H. Chen , M. Salazar-Palma and M-D Zhu, "Physics-Based Modeling of Experimental Data Encountered in Cellular Wireless Communication", *IEEE Transactions on Antennas and Propagation*, 2018 , Vol. 66 , Issue: 12, pp. 6673 – 6682.

APPENDIX A: DEFINITIONS OF THE VARIOUS WAVES USED IN THIS PAPER

In this appendix we present the definitions for the various types of waves that we have used in this paper.

Ground Wave: A wave from a source in the vicinity of a flat earth that would exist in space in the absence of the ionosphere. *NOTE*: A ground wave near the surface of the earth can be identified with a Norton surface wave for a grazing angle of incidence and when both the transmitter and the receiver are far away from each other on the earth's surface.

Lateral Wave: A wave guided along the interface between two media. For sufficiently large distances from the source the field decays as the power of 3/2 of the distance. This wave is usually present when a wave is incident from a denser to a rarer medium undergoing total internal reflection. In this case it displays the Goos-Hänchen effect of a lateral shift in the reflected wave. A lateral wave can be represented by a sum of leaky waves.

Norton surface wave: A propagating electromagnetic wave on the surface of the earth when both the transmitters and the receivers are close to the ground. Asymptotically this wave decays also as the square of the distance. It consists of the total field minus the geometric optics field and hence does not satisfy Maxwell's equations.

Space wave: A wave which propagates through the atmosphere, from a transmitter antenna to a receiving antenna. These waves can travel directly or can travel after reflecting from the earth's surface to the troposphere surface of earth.

Surface wave: A wave guided by a boundary of two dissimilar media and has a phase velocity smaller than the velocity of light. Its field perpendicular to the direction of propagation is evanescent in nature. As the frequency increases the wave is more confined to the interface. There are no radiative fields associated with a surface wave.

Zenneck wave: It is a wave that is a solution of Maxwell's equations and decays exponentially both in the transverse plane and along the direction of propagation. It has a phase velocity faster than the speed of light and the propagation constants of this wave are generally not highly dependent on frequency. In addition, a cylindrical Zenneck surface wave decays as $1/\sqrt{R}$. According to Schelkunoff this is strictly not a *true* surface wave.

Leaky wave: Leaky waves are solutions of the source-free Maxwell equations but they are not proper modes because of their singular behavior at infinity in the transverse, or crosssection, plane. Despite the decidedly unphysical nature of this aspect of the leaky-wave solutions, however, they are valid field representations in certain restricted regions within which they remain finite. It is the pur-pose of this section to expand on these remarks. A leaky wave travels along its guiding structure with a complex propagation constant; its phase velocity is greater than that of light and its attenuation constant is indicative of a continuous leakage of energy along the guiding structure. The improper transverse variation of the field of a leaky wave follows directly from the relationship among the wavenumbers for rectangular regions.

Although leaky waves are frequently assigned to the same class as surface waves, they are distinctly different in nature from true surface waves. A true surface wave, unlike a leaky wave, is a proper mode and possesses the following characteristics: a phase velocity ordinarily less than that of light, a real propagation wavenumber in the longitudinal direction, and a purely imaginary wavenumber in the transverse direction such that the wave decays away from the guiding structure. In contrast, a leaky wave is nonmodal and possesses a phase velocity along the surface greater than that of light, a complex propagation constant in the longitudinal direction, and a complex transverse wavenumber such that the wave propagates transversely away from the guiding structure with increasing amplitude. In addi-tion, these two wave types have different mechanisms of radiation. The surface wave is basically nonradiating and can produce radiation only at a discontinuity on the guiding structure, whereas the leaky wave radiates continuously.

APPENDIX B: ASYMPTOTIC EVALUATION OF THE INTEGRALS BY THE METHOD OF STEEPEST DESCENT

The method of steepest descent (or the saddle point method) deals with the approximate evaluation of integrals of the form

$$I(\rho)=\int_C F(\xi)\exp\left[-\rho f(\xi)\right]d\xi \tag{B.1}$$

for large values of ρ, where the contour C in the complex ξ plane is such that that integrand goes to zero at the ends of the contour. The functions $f(\xi)$ and $F(\xi)$ are arbitrary analytic functions of the complex variable ξ.

The basic philosophy of the method of steepest descent is as follows: A path is selected in the complex ξ plane in such a way that the entire value of the integral is determined from a comparatively short portion of the path. Within certain limits, the contour of integration C may be altered to such a path without affecting the value of the integral. Then the integral is replaced by another, simpler function, which closely approximates the integrand over the essential portions of the path. The behavior of the new integrand outside the important portion of the path is of no concern. For real and positive values of ρ and for a general contour C the quantity $\rho f(\xi)$ is positive on some parts of the path and there are other regions where it is negative. The latter regions are more important since the integrand is larger, and in those regions, where the negative of $\text{Re}\left[\rho f(\xi)\right]$ is the largest, it is important to reduce oscillations. A contour is chosen along which the imaginary part of $\left[\rho f(\xi)\right]$ is constant in the region where the negative of its real part is largest. The path in the region where $\text{Re}\left[\rho f(\xi)\right]$ is greatest may be chosen so that $\text{Im}\left[\rho f(\xi)\right]$ varies if this turns out to be necessary to complete the contour. In this way, the oscillations of the integral cause the least trouble. Since the path of integration must pass along the line of most rapid increase and decrease of $\text{Re}\left[f(\xi)\right]$ it must coincide with the line $\text{Im}\left[f(\xi)\right]$=constant, which may be a line of constant phase. The point of the path at which $\text{Re}\left[f(\xi)\right]$ is an extremum is called the saddle point and the derivative of $\text{Re}\left[f(\xi)\right]$ must be zero at this point. Since $\text{Im}\left[f(\xi)\right]$is a constant on this path, then its derivative must also be zero, and therefore

$$\frac{df}{d\xi}=0 \tag{B.2}$$

at the saddle point. Thus the most advantageous path of integration must go through the saddle point along the line of the most rapid decrease of the function $\mathrm{Re}\left[f\left(\xi\right)\right]$, which coincides with the line $\mathrm{Im}\left[f\left(\xi\right)\right]=$ constant. This path then is called the path of steepest descent. If the saddle point occurs at $\xi=\xi_0$, then it follows that the path of integration will be determined from

$$f\left(\xi\right)=f\left(\xi_0\right)+s^2 \tag{B.3}$$

where s is real and $-\infty\le s\le\infty$. The saddle point corresponds to the point $s=0$.

Now going back to the integral (B.1) and using (B.3)

$$I_{SD}=\exp\left[-\rho f\left(\xi_0\right)\right]\int_{-\infty}^{\infty}F\left(\xi\right)\exp\left[-\rho s^2\right]d\xi \tag{B.4}$$

If $\Phi\left(s\right)=F\left(\xi\right)\dfrac{d\xi}{ds}$ (B.5)

then $I_{SD}=\exp\left[-\rho f\left(\xi_0\right)\right]\displaystyle\int_{-\infty}^{\infty}\Phi\left(s\right)\exp\left[-\rho s^2\right]ds$ (B.6)

Now if ρ is large, then the integrand in (B.6) will fall off rapidly with an increasing value of s, the distance from the saddle point. Thus only small values of s will contribute significantly and, therefore, we can expand $\Phi\left(s\right)$ in a Taylor series about the saddle point $s=0$. Therefore we can write,

$$\Phi\left(s\right)=\Phi\left(0\right)+s\Phi'\left(0\right)+\frac{s^2}{2}\Phi''\left(0\right)+....$$

(B.7)

Substituting (B.7) into (B.6) one obtains

$$I_{SD}=\exp\left[-\rho f\left(\xi_0\right)\right]\int_{-\infty}^{\infty}\exp\left[-\rho s^2\right]\left\{\Phi\left(0\right)+\frac{s^2}{2}\Phi''\left(0\right)+..\right\}ds \tag{B.8}$$

as the odd powers of s will integrate to zero. Since

$$\int_{-\infty}^{\infty}\exp\left[-\rho s^2\right]s^{2n}ds=\frac{\sqrt{\pi}\left(2n\right)!}{n!\,2^{2n}}\frac{1}{\rho^{n+0.5}} \tag{B.9}$$

and substituting (B.9) into (B.8) yields

$$I_{SD}=\sqrt{\frac{\pi}{\rho}}\exp\left[-\rho f\left(\xi_0\right)\right]\left[\Phi\left(0\right)+\frac{1}{4\rho}\Phi''\left(0\right)+....\right] \tag{B.10}$$

Now we need to relate $\Phi\left(s\right)$ to $F\left(\xi\right)$ as described in (B.5). To make the connection, we first expand $f\left(\xi\right)$ in a Taylor series around the saddle point $f\left(\xi_0\right)$, and if $\xi-\xi_0=x$, then one obtains

$$f\left(\xi\right)=f\left(\xi_0\right)+\frac{x^2}{2!}f''\left(\xi_0\right)+\frac{x^3}{3!}f'''\left(\xi_0\right)+\frac{x^4}{4!}f^{IV}\left(\xi_0\right)+...=f\left(\xi_0\right)+s^2 \tag{B.11}$$

The goal here is to relate x, to a power series of s. To this end we get

$$x=a_0\,s\left(1+a_1s+a_2s^2+a_3s^3+....\right) \tag{B.12}$$

And therefore

$$x^2 = a_0{}^2 s^2\left[1+2a_1 s+\left(2a_2+a_1{}^2\right)s^2+....\right] \qquad \text{(B.13)}$$

$$x^3 = a_0{}^3 s^3\left[1+3a_1 s+3\left(a_1{}^2+a_2\right)s^2+....\right] \qquad \text{(B.14)}$$

$$x^4 = a_0{}^4 s^4\left[1+4a_1 s+2\left(2a_2+3a_1{}^2\right)s^2+....\right] \qquad \text{(B.15)}$$

Rewriting (B.11) we get

$$s^2 = Ax^2+Bx^3+Cx^4+... \qquad \text{(B.16)}$$

where

$$A=\frac{f''(\xi_0)}{2};\ B=\frac{f'''(\xi_0)}{6};\ C=\frac{f^{IV}(\xi_0)}{24} \qquad \text{(B.17)}$$

Now substituting (B.12)-(B.15) into (B.16), we get

$$s^2 = A\left[a_0{}^2s^2+2a_1a_0{}^2s^3+\left(a_1{}^2+2a_2\right)a_0{}^2s^4\right]+B\left[a_0{}^3s^3+3a_1a_0{}^3s^4+...\right]+C\left[a_0{}^4s^4+...\right] \qquad \text{(B.18)}$$

from which

$$a_0=\frac{1}{\sqrt{A}};\ a_1=-\frac{B}{2A^{3/2}};\ a_2=-\frac{C}{2A^2}+\frac{5}{8}\frac{B^2}{A^3} \qquad \text{(B.19)}$$

Next we expand the function $F(\xi)$ in a Taylor series around the saddle point ξ_0, and with $\xi-\xi_0=x$, we get

$$F(\xi)=F(\xi_0)\left[1+Px+Qx^2+...\right]$$

(B.20)

and $P=\dfrac{F'(\xi_0)}{F(\xi_0)}$ and $Q=\dfrac{F''(\xi_0)}{2F(\xi_0)}$ … (B.21)

Using (B.11), (B.13)-(B.15), (B.19), and (B.20), we get

$$\begin{aligned}
\Phi(s) &= F(\xi)\frac{d\xi}{ds} = F(\xi)\frac{dx}{ds}\\
&= F(\xi)\left[a_0\left(1+a_1s+a_1s^2+..\right)+a_0\,s\left(a_1+2a_2s+....\right)+\right]\\
&= F(\xi)\left[a_0+s\left(2a_0a_1\right)+s^2\left(a_0a_2{}^2+2a_2a_0\right)+...\right]\\
&= F(\xi_0)\left[1+Px+Qx^2+...\right]\left[a_0+2a_0a_1s+s^2\left(a_0a_2{}^2+2a_2a_0\right)+...\right]\\
&= F(\xi_0)\left[1+a_0Ps+a_0a_1Ps^2+a_0a_2Ps^3+a_0{}^2Qs^2+2a_0{}^2a_1Qs^3+...\right]\left[a_0+2a_0a_1s+3a_0a_2s^2+...\right] \qquad \text{(B.22)}\\
&= F(\xi_0)\left[a_0+s\left(a_0{}^2P+2a_0a_1\right)+s^2\left(a_0{}^2a_1P+2a_0{}^2a_1P+3a_0a_2+a_0{}^3Q\right)+....\right]\\
&= \frac{F(\xi_0)}{\sqrt{A}}\left[1+s\left(\frac{P}{\sqrt{A}}-\frac{B}{A^{3/2}}\right)+s^2\left(\frac{Q}{A}-3\left\{\frac{C}{2A^2}-\frac{5B^2}{8A^3}\right\}-\frac{3}{\sqrt{A}}\frac{BP}{2A^{3/2}}\right)+...\right]\\
&= \frac{F(\xi_0)}{\sqrt{A}}\left[1+s\left(\frac{P}{\sqrt{A}}-\frac{B}{A^{3/2}}\right)+s^2\left(\frac{Q}{A}+\frac{15B^2}{8A^3}-\frac{3BP}{2A^2}-\frac{3C}{2A^2}\right)+...\right]
\end{aligned}$$

Therefore

$$\Phi(0)=\sqrt{\frac{2}{f^{II}(\xi_0)}}\;F(\xi_0) \tag{B.23}$$

$$\Phi^{II}(0)=2\Phi(0)\left[\frac{F^{II}}{Ff^{II}}+\frac{5\left(f^{III}\right)^2}{12(f^{II})^3}-\frac{f^{III}}{\left(f^{II}\right)^2}\frac{F^{I}}{F}-\frac{f^{iv}}{4\left(f^{II}\right)^2}\right] \tag{B.24}$$

Hence

$$I_{sd}=\sqrt{\frac{2\pi}{\rho f^{II}(\xi_0)}}\,F(\xi_0)\exp\left[-\rho f(\xi_0)\right]\left\{1+\frac{1}{2\rho}\left[\frac{F^{II}}{Ff^{II}}+\frac{5\left(f^{III}\right)^2}{12\left(f^{II}\right)^3}-\frac{f^{III}}{\left(f^{II}\right)^2}\frac{F^{I}}{F}-\frac{f^{IV}}{4\left(f^{II}\right)^2}\right]\right\}+\ldots \tag{B.25}$$

The interesting point regarding the result of (B.25) is that it is a divergent series for a fixed ρ as the number of terms in this expansion increases for a fixed value of ρ. Such a series is called an asymptotic series as introduced by Poincaré.

A divergent series is one

$$g(\rho)=A_0+\frac{A_1}{\rho}+\frac{A_2}{\rho^2}+\ldots+\frac{A_n}{\rho^n} \tag{B.26}$$

in which the sum of the first $(n+1)$ terms is $S_n(z)$, given by

$$S_n(\rho)=\sum_{L=0}^{n}A_i\rho^{-i} \tag{B.27}$$

which is said to be an asymptotic expansion of a function $g(\rho)$ for a given range of argument ρ if the expression

$$R_n(\rho)=\rho^n\left\{g(\rho)-S_n(\rho)\right\} \tag{B.28}$$

satisfies the condition

$$\lim_{|\rho|\to\infty}R_n(\rho)=0 \text{ for } n \text{ fixed} \tag{B.29}$$

even though

$$\lim_{n\to\infty}\left|R_n(\rho)\right|=\infty,\ \rho-\text{fixed.} \tag{B.30}$$

When this is the case, one can make

$$\left|R_n(\rho)\right|=\left|\rho^n\left\{g(\rho)-S_n(\rho)\right\}\right|<\varepsilon \tag{B.31}$$

where ε is arbitrarily small, by making $|\rho|$ sufficiently large. Some of the properties of this definition are:

(a) Asymptotic expansions can be multiplied unconditionally.
(b) Asymptotic expansions can be integrated unconditionally.
(c) An asymptotic expansion of a function is unique.
(d) One asymptotic expansion may represent several functions.
(e) Asymptotic expansions can be divided providing the divisor contains at least one non-zero coefficient.

The point about the series of (B.25) is that for sufficiently large values of $|\rho|$ the terms of the series decrease at least initially, and that if the series is truncated before the smallest term, the error is of the order of magnitude of the first discarded term. So, if $\Phi(s)$ is any function in (B.6) for which $\int_{-\infty}^{\infty} \Phi(s)\exp\left[-\rho s^2\right] ds$ converges for sufficiently large values of the parameter ρ, then the asymptotic expansion of (B.6) in descending powers of $\sqrt{\rho}$ can be given by replacing $\Phi(s)$ by a Taylor series in ascending powers of s in (B.7) and then integrating term by term. In that case (B.27) is the asymptotic expansion of (B.6). In terms of the problem for our case, the integrals that we will be dealing with are of the following form.

$$I(kR) = \int F(\beta)\, \exp\left[-j\,k\,R\cos(\beta-\theta)\right] d\beta \qquad \text{(B.31)}$$

so that $\rho = j\,k\,R$

(B.32)

$$f(\beta) = \cos(\beta-\theta) \qquad \text{(B.33)}$$

The saddle point then occurs at $\beta = 0$ and we get

$f(\theta)=1,\ f^{I}(\theta)=0,\ f^{II}(\theta)=-1,\ f^{III}(\theta)=0,\ f^{IV}(\theta)=1$.

Substituting these values in (A.25) we get

$$I_{SD}(kR) = \sqrt{\frac{2\pi j}{kR}}\, F(\theta)\, e^{-jkR}\left\{1+\frac{j}{2kR}\left(\frac{F^{II}(\vartheta)}{F(\vartheta)}+\frac{1}{4}\right)\right\}+\ldots \qquad \text{(B.34)}$$

APPENDIX C: ASYMPTOTIC EVALUATION OF THE INTEGRALS WHEN THERE EXISTS A POLE NEAR THE SADDLE POINT

The asymptotic expansion given by (A.34) is not valid if there is a pole near the saddle point θ. However, the method of steepest descent can be modified in such a way that the presence of poles is taken into account from the very beginning in evaluating these integrals. Of special interest in the analysis will be an integral of the form

$$I(kR) = \int_{\Gamma_1} F_1(\beta)\exp\left[-jkR\cos(\beta-\theta)\, d\beta\right. \qquad \text{(C.1)}$$

where Γ_1 is a path of integration in the complex β plane as discussed in the previous sections. $F_1(\beta)$ now has a pole β_P near the saddle point θ. For large values of kR, the pole can be factored out from $F_1(\beta)$ by writing $F_1(\beta) = \dfrac{F(\beta)}{\sin\left(\dfrac{\beta-\beta_P}{2}\right)}$. It is then argued that since $F(\beta)$ has no singularities in the vicinity of the saddle point, it may be removed from under the integral sign with β equated to θ, as presented by Clemmow [32]. Thus the integral of (C.1) can be written as

$$I_{SD}(kR) = F(\theta)\int_{\Gamma_1} \frac{\exp\left[-j\,k\,R\cos(\beta-\theta)\right]}{\sin\left(\dfrac{\beta-\beta_P}{2}\right)} d\beta \qquad \text{(C.2)}$$

$$= F(\theta)\int_{\Gamma_0} \frac{\exp\left[-j\,k\,R\cos\alpha\right]}{\sin\left(\dfrac{\alpha+\theta-\beta_P}{2}\right)} d\alpha \qquad \text{(C.3)}$$

where $\alpha = \beta - \theta$. By reversing the sign of α as

$$I_{SD}\left(kR\right) = F\left(\theta\right)\int_{\Gamma_0} \frac{\exp\left[-j\,k\,R\cos\alpha\right]}{\sin\left(\dfrac{\theta-\alpha-\beta_P}{2}\right)}d\alpha \tag{C.4}$$

and then adding (C.3) to (C.4) and then dividing by two will convert (C.2) to

$$I_{SD}\left(kR\right) = 2\sin\frac{\gamma}{2}F\left(\theta\right)\int_{\Gamma_0} \frac{\exp\left[-j\,k\,R\cos\,\alpha\right]}{\cos\,\alpha-\cos\,\gamma}\cos\frac{\alpha}{2}\,d\alpha \tag{C.5}$$

where $\gamma = \theta - \beta_P$. Now by changing the variable of integration from α to τ such that

$$\tau = \sqrt{2}\exp\left[-j\,{}^{\pi}\!/_{4}\right]\sin\frac{\alpha}{2} \tag{C.6}$$

the path Γ_0 is now transformed to an integral from $-\infty$ to $+\infty$. Hence

$$I_{SD}\left(kR\right) = 2b\exp\left[-j\,k\,R + j\,{}^{3\pi}\!/_{4}\right]F\left(\theta\right)\int_{-\infty}^{\infty}\frac{e^{-k\,R\,\tau^2}}{\tau^2+jb^2}\,d\tau \tag{C.7}$$

where

$$b = \sqrt{2}\sin\frac{\gamma}{2} \tag{C.8}$$

since

$$\int_{-\infty}^{\infty}\frac{e^{-k\,R\,\tau^2}}{\tau^2+jb^2}d\tau = \frac{\pi}{b}\exp\left[j\left(b^2k\,R-{}^{\pi}\!/_{4}\right)\right]\times\text{erfc}\left(\sqrt{j\,k\,R\,b^2}\right) \tag{C.9}$$

and

$$W^2 = -\,j\,k\,R\,b^2 = -j\,2\,k\,R\sin^2\left(\frac{\theta-\beta_P}{2}\right) \tag{C.10}$$

then (C.7) becomes

$$I_{SD}\left(k\,R\right) = 2\pi\,j\,F(\theta)\exp\left[-j\,k\,R-W^2\right]\,\text{erfc}\left(jW\right). \tag{C.11}$$

This completes the derivation.