ELSEVIER

Signal Processing 79 (1999) 301–307

SIGNAL PROCESSING

www.elsevier.nl/locate/sigpro

On an ambiguity in the definition of the amplitude and phase of a signal

Leon Cohen[a,*,1], Patrick Loughlin[b,2], David Vakman[c]

[a]*City University of New York, Hunter College and Graduate Center, 695 Park Avenue, New York, NY 10021, USA*
[b]*University of Pittsburgh, 348 Benedum Hall, Pittsburgh, PA 15261, USA*
[c]*213 Bennet Avenue, New York, NY 10040, USA*

Received 8 February 1999; received in revised form 29 June 1999

Abstract

We point out that the conventional definition of instantaneous amplitude and frequency, namely as the magnitude and derivative of the phase, respectively, of a complex representation of the signal, sometimes contains an ambiguity, even for a unique complex representation (e.g., the analytic signal). There are at least two choices for resolving this ambiguity when it arises. One choice yields a nonnegative amplitude but an instantaneous frequency with infinite spikes, and one yields a bounded instantaneous frequency but an instantaneous amplitude with positive and negative values. Historically, both solutions (i.e., both amplitudes) have been important in radio engineering, and both can be measured with real devices. The former choice is more commonly used for defining the instantaneous amplitude and frequency of signals, but the latter choice is equally acceptable and may be preferred in some situations.

Zusammenfassung

Wir machen darauf aufmerksam, daß die herkömmliche Definition der Momentanamplitude und -frequenz, nämlich als Betrag respektive Ableitung der Phase einer komplexen Darstellung des Signals, manchmal eine Mehrdeutigkeit aufweist; dies gilt sogar für eine eindeutige komplexe Darstellung (z.B. das analytische Signal). Es gibt zumindest zwei Möglichkeiten, diese Mehrdeutigkeit zu beseitigen. Eine Möglichkeit führt auf eine nichtnegative Amplitude, aber andererseits auf eine Momentanfrequenz mit unendlichen Spitzen; eine andere führt auf eine beschränkte Momentanfrequenz, aber andererseits auf eine Momentanamplitude mit positiven und negativen Werten. Historisch gesehen waren beide Lösungen (d.h. beide Amplituden) in der Radiotechnik von Bedeutung, und beide können mit realen Geräten gemessen werden. Die erste Möglichkeit wird häufiger verwendet, um die Momentanamplitude und -frequenz von Signalen zu definieren, aber die zweite Möglichkeit ist genauso akzeptabel und könnte in manchen Situationen bevorzugt werden.

* Corresponding author. Tel.: + 1-212-772-5248.
E-mail address: lcchc@cunyvm.cuny.edu (L. Cohen)

[1] Work supported by the Office of Naval Research, the NASA JOVE and the NASA HBCU/MI programs.
[2] Work supported by the National Science Foundation (grant no. MIP-9624089)

0165-1684/99/$ - see front matter © 1999 Elsevier Science B.V. All rights reserved.
PII: S0165-1684(99)00103-6

Résumé

Nous mettons en lumière le fait que la définition conventionnelle de l'amplitude et la fréquence instantanées en tant qu'amplitude et dérivée de la phase, respectivement, d'une représentation complexe du signal, donne parfois lieu à une ambiguïté, même pour une représentation complex unique (à savoir le signal analytique). Il existe au moins deux choix pour résoudre cette ambiguïté lorsqu'elle se présente. Un choix fournit une amplitude non négative mais une fréquence instantanée avec des pics infinis, et l'autre une fréquence instantanée bornée mais une amplitude instantanée ayant des valeurs positives et négatives. Historiquement, les deux solutions (les deux amplitudes) ont été importantes en radio, et toutes deux peuvent être mesurées avec des appareils réels. Le premier choix est plus communément utilisé pour définir l'amplitude et la fréquence instantanées des signaux, mais le deuxième est également acceptable et peut être préférable dans certaines situations.

Keywords: Instantaneous frequency; Analytic signal; Hilbert transform

1. Introduction

The main motivation of Gabor for defining the analytic signal was to provide an unambiguous definition of instantaneous frequency [1]. Gabor (and later Ville [6]) argued that to define instantaneous frequency we have to first define a complex signal from which the instantaneous frequency is then the derivative of the phase. But given a real signal, there is an infinite number of complex signals whose real part is the given real signal, but whose imaginary parts – and therefore instantaneous amplitude, phase and frequency (APF) – are different. Gabor proposed a method for choosing a particular APF for a given real signal. Gabor's prescription was to "suppress the amplitudes belonging to negative frequencies (in the spectrum of the real signal), and multiply the amplitudes of positive frequencies by two [1]." Inverse Fourier transformation of this modified spectrum yields a complex signal, namely the analytic signal, whose real part equals the given real signal and whose imaginary part equals the Hilbert transform of the real signal. It has since been shown that the analytic signal can be derived from just a few reasonable physical conditions [5]. We discuss these conditions in Section 3.

Having thus defined a particular complex signal, we can express it as amplitude and phase in the usual way,

$$z(t) = x(t) + \mathrm{j}y(t) = A(t)\,\mathrm{e}^{\mathrm{j}\phi(t)}, \tag{1.1}$$

where the amplitude and phase are commonly given as

$$A(t) = \sqrt{x^2(t) + y^2(t)}, \quad \phi(t) = \arctan y(t)/x(t), \tag{1.2}$$

and the instantaneous frequency is

$$\omega_i(t) = \phi'(t) = (y'x - x'y)/A^2. \tag{1.3}$$

Eq. (1.2) is the one that is generally written in most articles dealing with these issues for the amplitude and phase of the complex representation. It is the purpose of this note to point out that sometimes there is an ambiguity that arises in specifying the amplitude and phase, *even when a unique real and imaginary part is given*. To show when this ambiguity arises, we begin with a simple example to motivate and develop the main ideas. We remark that our considerations are not restricted to the analytic signal, but apply to complex representations in general.

2. Example

Consider the signal

$$s(t) = \tfrac{1}{2}\cos\omega_a t + \tfrac{1}{2}\cos w_b t = \cos\omega_1 t\cos\omega_2 t, \tag{2.1}$$

where

$$\omega_a = \omega_2 + \omega_1, \quad \omega_b = \omega_2 - \omega_1. \tag{2.2}$$

For $\omega_2 > \omega_1 \geqslant 0$, the corresponding analytic signal is

$$\begin{aligned} z(t) &= \cos\omega_1 t\cos\omega_2 t + \mathrm{j}\cos\omega_1 t\sin\omega_2 t \\ &= \cos(\omega_1 t)\,\mathrm{e}^{\mathrm{j}\omega_2 t}. \end{aligned} \tag{2.3}$$

What are the amplitude and phase? According to the usual procedure, they are

$$A(t) = \sqrt{\cos^2\omega_1 t \cos^2\omega_2 t + \cos^2\omega_1 t \sin^2\omega_2 t}$$
$$= |\cos \omega_1 t| \tag{2.4}$$

and

$$\phi(t) = \arctan[\cos \omega_1 t \sin \omega_2 t/\cos \omega_1 t \cos \omega_2 t]$$
$$= \omega_2 t, \tag{2.5}$$

from which we then have

$$z(t) = |\cos \omega_1 \ t| e^{j\omega_2 t}. \tag{2.6}$$

But that is not correct! Specifically, the real part is not the given signal,

$$\mathscr{R}\{|\cos \omega_1 t| e^{jw_2 t}\}$$
$$= |\cos \omega_1 t| \cos \omega_2 t \neq \cos \omega_1 t \cos \omega_2 t. \tag{2.7}$$

The correct amplitude-phase is

$$z(t) = \cos(\omega_1 t)\, e^{j\omega_2 t}. \tag{2.8}$$

Note that the instantaneous amplitude for this signal is not strictly nonnegative, yet by convention instantaneous amplitude is almost universally taken to be (or assumed) nonnegative. Technically, $A(t) = \sqrt{x^2(t) + y^2(t)}$ is more correctly the magnitude of the complex signal, and we make that point explicit in this paper by writing $|A(t)| = \sqrt{x^2(t) + y^2(t)}$. We may of course insist that the instantaneous amplitude equals the magnitude, but the phase must then be modified in order that the correct real part is obtained. Specifically, for this signal, if we take $A(t) = |A(t)|$, then the phase must incur discontinuities and the correct 'magnitude-phase' representation of this signal is

$$z(t) = |\cos \omega_1 t| e^{j[\omega_2 t + \frac{1}{2}\pi(1 - \mathrm{sgn}(\cos\ \omega_1 t))]}. \tag{2.9}$$

The conflict can thus be eliminated when we take $A(t) = |A(t)|$ by using the fact that the inverse tangent is not a single-valued function: $\arctan(\tan \omega_2 t) = \omega_2 t + k\pi$. *However*, we note that here k is not a simple constant (integer), but rather a function of time because we have a choice *at each instant of time* when evaluating the arctangent. In particular, for the example above,

$$k = k(t) = \tfrac{1}{2}(1 - \mathrm{sgn}(\cos \omega_1 t)). \tag{2.10}$$

The two expressions above for the amplitude and phase are clearly quite different, yet both expressions yield the identical complex signal, $\cos \omega_1 t \cos \omega_2 t + j \cos \omega_1 t \sin \omega_2 t$. Hence, the notion that specifying a unique complex representation fixes the amplitude and phase of a signal is in fact not always true: there are situations where an infinite number of possibilities for the amplitude and phase, given a unique complex representation, arise. We have given two possibilities above for a particular signal, but there are actually an unlimited number for that signal. Consider, for example, the infinity of possibilities for the amplitude that lie between full rectification $|\cos \omega_1 t|$ to no rectification $\cos \omega_1 t$: we may rectify $\cos \omega_1 t$ over any period(s), and compensate with an appropriate jump in the corresponding phase (see Fig. 1). Is this situation a rarity, unique to this particular example? Finally, what are the consequences of insisting that instantaneous amplitude is always nonnegative, i.e., in taking $A(t) = \sqrt{x^2(t) + y^2(t)}$? We explore these questions next.

3. The general problem

To begin, let us consider briefly the conditions given in [5] for the amplitude and phase of a signal. Given a real signal $x(t)$, we desire to associate with it another signal $y(t) = H[x(t)]$, where H is as yet some undefined operator, from which we obtain the complex signal $z(t) = x(t) + jH[x(t)]$ with amplitude and phase as given by Eq. (1.2). What is the operator H? Vakman proposed that the amplitude and phase it generates should have the following three physical properties:

1. *Amplitude continuity*: a small change in the value of the signal $x(t)$ should induce a correspondingly small change in the instantaneous amplitude $A(t)$.
2. *Phase independence of scale*: multiplying the real signal $x(t)$ by a real positive constant c should have no affect on the instantaneous phase and

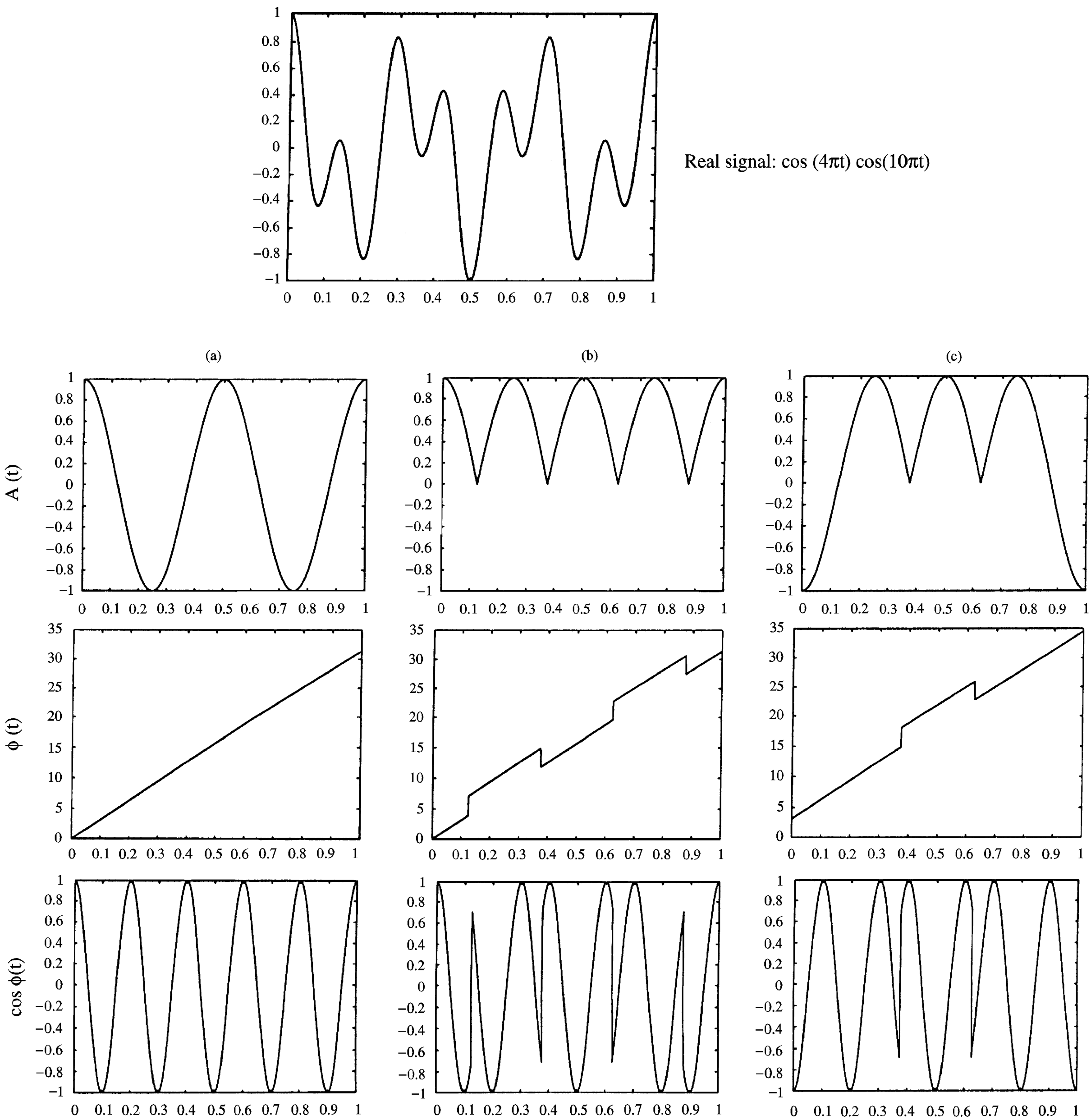

Fig. 1. Three possible different amplitude-phase pairs (a, b, c), as obtained via the Hilbert transform, for the real signal at top. For each case, we show (top to bottom) the amplitude, $A(t)$, phase, $\phi(t)$, and the cosine of the phase, $\cos(\phi(t))$. The complex (and analytic) signal for each case is $A(t)e^{j\phi(t)}$, and the real signal is $A(t)\cos\phi(t)$. Note that although the amplitudes and phases are different in each case, they produce the *identical* real and complex signals. This illustrates that even when the complex form is unique, there can be an ambiguity in specifying the amplitude and phase. For this case we may rectify the amplitude $A(t)$ in (a) over any period(s) and compensate with a jump of π radians in the corresponding phase. Note that although all amplitudes are continuous, the phase is continuous only in case (a).

frequency and should multiply the instantaneous amplitude by the same constant.

3. *Harmonic correspondence*: the instantaneous amplitude, phase and frequency of a pure sinusoid $A_0 \cos(\omega_0 t + \phi_0)$ should be given, respectively, by $A(t) = A_0$, $\phi(t) = \omega_0 t + \phi_0$, and $\omega_i(t) = \phi'(t) = \omega_0$.

Remarkably, these three simple, reasonable conditions force the operator H to be the Hilbert transform, meaning that the APF satisfying these conditions are those obtained from the analytic signal representation of the given real signal.[3]

With these considerations in mind, we are now in a position to consider when the ambiguity pointed out previously arises. We write the complex signal in two equivalent forms,

$$z(t) = A(t)e^{j\phi(t)} \tag{3.1}$$

$$= |A(t)|e^{j\phi(t) + j\pi\alpha(t)}, \tag{3.2}$$

where

$$A(t) = z(t)e^{j\phi(t)} = x(t)\cos\phi(t) + y(t)\sin\phi(t) \tag{3.3}$$

is not restricted to be nonnegative, and where $\alpha(t)$ is a function of time whose value can only be 1 or 0. We note again that there is actually an infinite number of equivalent ways to express the *same* complex representation, ranging between full rectification of the amplitude as in Eq. (3.2) to no rectification as in Eq. (3.1).[4] However, these two extremes are sufficient to address the general problem, to show when it arises, and to suggest possible resolutions to this ambiguity.

Resolution 1. Always take the amplitude to be nonnegative, i.e., $A(t) = |z(t)| = \sqrt{x^2(t) + y^2(t)}$. If we then require amplitude continuity, the ambiguity occurs only when the magnitude of the complex representation, $|z(t)|$, is not positive-definite (i.e., when it is zero at certain instants of time), since it is at those times that the amplitude may or may not cross zero; for example, is it $|\cos\omega_1 t|$ or $\cos\omega_1 t$? At those time instants, the phase may incur a discontinuity, i.e., $\alpha(t)$ will jump from 0 to 1 or 1 to 0. However, there is no ambiguity when the magnitude of the complex representation is positive-definite, since then we have $|z(t)| = |A(t)| > 0$, which yields, by amplitude continuity, $A(t) = |A(t)| = \sqrt{x^2 + y^2}$, and no compensation in the phase is necessary.

Resolution 2. Another possibility is to insist that, given a continuous, differentiable real function, the amplitude *and phase* both be continuous. Indeed, for the example given previously, this constraint eliminates all but the choice $\cos\omega_1 t$ from the set of possibilities ranging from $|\cos\omega_1 t|$ to $\cos\omega_1 t$ for the amplitude of the signal above, because only that choice yields a continuous phase.

Hence, while the conditions in [5] determine a unique operator H for determining the instantaneous amplitude, phase and frequency of a signal, they do not in fact eliminate the ambiguity that arises in determining the amplitude from the complex form for situations when the magnitude is zero at certain times. The *phase continuity* condition given here resolves this ambiguity. Insistence on a nonnegative amplitude also resolves the ambiguity but yields infinite instantaneous frequency at zero magnitude when the phase jumps by π.

3.1. Limiting procedure

Some further insight can be gained if we consider two tones of unequal amplitude and we consider what happens to the phase as the amplitudes of each tone become equal.

For that case, the order in which one evaluates the equation makes a difference, because the amplitude is zero at specific times for equal strength tones. Consider

$$x(t) = A_a \cos\omega_a t + A_b \cos\omega_b t, \tag{3.4}$$

with $A_a \geqslant A_b > 0$. The instantaneous frequency is

$$\phi'(t) = \tfrac{1}{2}(\omega_a + \omega_b) + \tfrac{1}{2}(\omega_a - \omega_b)\frac{A_a^2 - A_b^2}{A^2(t)}, \tag{3.5}$$

[3] We note that other physical conditions lead to other operators, and hence complex signals that are not analytic [2,3].

[4] As an example see Fig. 1, and in particular the amplitude-phase pair corresponding to case (c).

with

$$A^2(t) = A_a^2 + A_b^2 + 2A_aA_b\cos(\omega_a - \omega_b)t. \tag{3.6}$$

Note that $A^2(t) = 0$ periodically for equal strength components. At times when $\cos(\omega_a - \omega_b)t = -1$, we have [4]

$$\phi'(t)_{t=\pi(1+2n)/(\omega_a-\omega_b)} = \tfrac{1}{2}(\omega_a + \omega_b) + \tfrac{1}{2}(\omega_a - \omega_b)\frac{A_a + A_b}{A_a - A_b}. \tag{3.7}$$

Taking the limit $A_a \to A_b$ of this expression yields an unbounded instantaneous frequency at these times. Thus, evaluating the expression this way, we have that for equal strength tones, the instantaneous frequency is infinite at the times when the amplitude is zero:

$$\phi'(t) = \begin{cases} \tfrac{1}{2}(\omega_a + \omega_b) & \text{for } t \neq \pi(1+2n)/(\omega_a - \omega_b), \\ \infty & t = \pi(1+2n)/(\omega_a - \omega_b). \end{cases} \tag{3.8}$$

Apparently this result is physically difficult to reconcile with the fact that the complex signal is continuous and given by $\cos\omega_1 t\, e^{j\omega_2 t}$. If instead we first take $A_a \to A_b$ in Eq. (3.5), then the numerator of the second term is always zero and the instantaneous frequency is $\phi'(t) = \frac{1}{2}(\omega_a + \omega_b) = \omega_2$ for all times – which is a more physically satisfying answer. Both, however, are mathematically correct, with the former corresponding to an instantaneous amplitude of $|\cos\omega_1 t|$ and the latter corresponding to $\cos\omega_1 t$. We now show that the instantaneous amplitude corresponding to both procedures above can be physically realized.

4. Measuring instantaneous amplitude

For the specific case considered above, one can ask which of the two amplitudes $|\cos\omega_1 t|$ or $\cos\omega_1 t$, is the "correct" one? We point out that both choices have been profitably used in well-known practical devices that measure both amplitudes.

The first is the linear detector that contains a rectifier and a low-pass filter. For the real signal Eq. (2.1) with high frequency $\omega_2 \gg \omega_1$, it produces

$$f(|\cos\omega_1 t\cos\omega_2 t|) = f(|\cos\omega_1 t|\,|\cos\omega_2 t|) \approx c\,|\cos\omega_1 t|, \tag{4.1}$$

where c is the average of $|\cos\omega_2 t|$ obtained after low-pass filtering with the filter $f(\,)$. Thus, the linear detector device measures $|A(t)| = |\cos\omega_1 t|$ which is commonly called the amplitude in radio-engineering.

A second device is the synchronous detector. One multiplies the signal by $\cos\omega_2 t$; a low-pass filter is then applied which produces

$$f(\cos\omega_1 t\cos^2\omega_2 t) = f\left(\cos\omega_1 t\left[\tfrac{1}{2} + \tfrac{1}{2}\cos 2\omega_2 t\right]\right) \approx \tfrac{1}{2}\cos\omega_1 t \tag{4.2}$$

after filtering. Thus, by this device we have $A(t) = \cos\omega_1 t$. In communications engineering, this operation is often said to give the *in-phase* signal component. The quadrature component is obtained by multiplying by $\sin\omega_2 t$ instead, and then filtering (which for this signal, yields zero). The square root of the in-phase component squared plus the quadrature component squared is sometimes referred to as the signal amplitude or *envelope* (which here is $|A(t)|$).

Hence both representations have been used and have played important roles in radio-engineering, and either one can be measured by real devices.

5. Conclusion

We have pointed out that specifying a unique complex representation for a signal does not necessarily uniquely determine the amplitude and phase of the signal. The ambiguity arises (for continuous amplitudes) when the magnitude of the complex representation is zero at particular instants of time; at those times, the amplitude may or may not cross through zero, and the phase may or may not incur discontinuities, yielding infinite spikes in the instantaneous frequency. Mathematically, any of these possibilities is legitimate (since all of the corresponding amplitude-phase pairs yield the same complex signal), and thus the ambiguity must be resolved either by choosing one mathematical convention, or based on physical considerations.

If a real signal is well behaved, that is, it has no discontinuities and is differentiable, then we may expect that the associated complex signal should produce physical quantities that are also well behaved, that is, have no peculiarities such as discontinuities and singularities. Hence, if one considers instantaneous frequency as a physical quantity, then it is reasonable to expect it to be well behaved (e.g., not ranging to infinity) for real signals that have no peculiarities. In addition, we should expect that the amplitude be well behaved and in particular it should be continuous. This reasoning supports the argument that for differentiable real signals we should define not only the instantaneous amplitude to be continuous (the first condition in [5]) but also the phase should be continuous so that there are no singularities in the instantaneous frequency, and so that the ambiguity noted in this paper is resolved. However, since both definitions of amplitude, namely

$$A(t) = \sqrt{x^2(t) + y^2(t)} \qquad \text{Definition1}, \quad (5.1)$$

$$A(t) = x(t)\cos\phi(t) + y(t)\sin\phi(t) \quad \text{Definition2} \quad (5.2)$$

can be measured with real devices and have been used in practice, and the ambiguity noted can also be resolved by taking a discontinuous phase corresponding to the nonnegative continuous amplitude, the choice is then a matter of taste and convenience.

References

[1] D. Gabor, Theory of communication, IEE J. Comm. Eng. 93 (1946) 429–457.

[2] P. Loughlin, Do bounded signals have bounded amplitudes?, Multi. Systems Signal Process. 9 (4) (1998) 419–424.

[3] P. Loughlin, B. Tacer, On the amplitude- and frequency-modulation decomposition of signals, J. Acoust. Soc. Amer. 100 (3) (1996) 1594–1601.

[4] L. Mandel, Interpretation of instantaneous frequencies, Amer. J. Phys. 42 (1974) 840–846.

[5] D. Vakman, On the analytic signal, the Teager–Kaiser energy algorithm, and other methods for defining amplitude and frequency, IEEE Trans. Signal Process. 44 (4) (1996) 791–797.

[6] J. Ville, Theory and Applications of the Notion of Complex Signal, translated by I. Seline in RAND Tech. Rpt. T-92, RAND Corp., Santa Monica, CA, 1958.