Instantaneous frequency decomposition: An application to spectrally sparse sounds with fast frequency modulations

T. J. Gardner a) and M. O. Magnasco

Laboratory of Mathematical Physics, The Rockefeller University, 1230 York Ave, New York, New York 10021

(Received July 2002; Revised 6 January 2005; accepted 11 January 2005)

Classical time–frequency analysis is based on the amplitude responses of bandpass filters, discarding phase information. Instantaneous frequency analysis, in contrast, is based on the derivatives of these phases. This method of frequency calculation is of interest for its high precision and for reasons of similarity to cochlear encoding of sound. This article describes a methodology for high resolution analysis of sparse sounds, based on instantaneous frequencies. In this method, a comparison between tonotopic and instantaneous frequency information is introduced to select filter positions that are well matched to the signal. Second, a cross-check that compares frequency estimates from neighboring channels is used to optimize filter bandwidth, and to signal the quality of the estimate. These cross-checks lead to an optimal time–frequency representation without requiring any prior information about the signal. When applied to a signal that is sufficiently sparse, the method decomposes the signal into separate time–frequency contours that are tracked with high precision. Alternatively, if the signal is spectrally too dense, neighboring channels generate inconsistent estimates—a feature that allows the method to assess its own validity in particular contexts. Similar optimization principles may be present in cochlear encoding. © 2005 Acoustical Society of America. [DOI: 10.1121/1.1863072]

PACS numbers: 43.60.Ac, 43.60.Hj, 43.58.Ta, 43.64.Bt [ADP]

Pages: 2896–2903

I. INTRODUCTION

Time–frequency analysis is a general methodology for representing sound in two dimensions, time and frequency. This is an intuitive representation, evinced by the evolution of the musical score, which since ancient times has shown time horizontally and pitch vertically. Time–frequency analysis is limited by the uncertainty principle: the resolution of frequency measurements is inversely proportional to the resolution of temporal measurements, so the time–frequency plane has a fundamental “granularity.” However, while this limit holds for signals drawn from arbitrary ensembles, special classes of signals may have features permitting a higher resolution analysis.

Many methods exist for the analysis of sparse signals, i.e., those composed of a number of well-separated tones with limited amplitude and frequency modulation rates. For example, Greenewalt employed periodicity analysis to great success in his classic study of the acoustics of bird song. One family of methodologies for the analysis of sparse signals is based on the calculation of instantaneous frequencies—the phase derivatives of a complex filter bank. Though these methods are capable of representing sparse signals with high precision, they require prior information about the analyzed signal to choose the positions and bandwidth of the filters that contribute to the analysis. A general method for optimizing these parameters remains an open problem.

Instantaneous frequency decomposition (IFD) provides a methodology for optimizing the parameters of an instantaneous frequency analysis, without reference to any prior information about the analyzed signal. The method consists of two phases: an expansive phase in which the input signal is split through bandpass filtering into a highly redundant array of channels, and a contractive phase, in which the redundant channels are checked for agreement, or “consensus” and collapsed back together. Consensus between neighboring channels indicates the quality of the local frequency estimates, and is used to guide optimization of filter bandwidths. If the signal is sufficiently sparse, the time–frequency representation generated by the IFD will track the individual components of the signal with high precision. If not, poor consensus measures signal the failure of the method.

While our purpose in this article is to describe a practical tool for the high-precision analysis of sparse sounds, it is worthwhile to note its biological motivation. In one of the earliest views of cochlear function, frequency is determined by the spatial, or tonotopic, position of active auditory nerve fibers. An alternative form of frequency coding can be found in the phase-locked responses of auditory hair cells for frequencies below 4 kHz, auditory nerve fibers preferentially initiate action potentials at particular phases of the driving force. Licklider in 1951 suggested that the intervals between phase-locked spikes leads to a second representation of frequency that is independent of the spatial arrangement of auditory fibers. This representation of sound has been experimentally and conceptually supported through neurophysiology, psychophysics, and functional brain imaging.

The method of instantaneous frequency decomposition is conceptually related to this spike-interval based coding in the auditory nerve, and provides a rationale for combining tonotopic and phase information in a single analysis, and for comparing frequency estimates from a re-
II. METHOD
A. Definitions

The continuous Gabor transform, also known as the short-time Fourier transform, is defined in terms of the signal to be analyzed $s$, a windowing function $w$, time $t$, and frequency $f^1$:

$$G_w(t,f) = \int s(\tau)w(\tau-t)e^{i2\pi f(\tau-t)}d\tau.$$

(1)

Gaussian windows are used throughout this article:

$$w = e^{-(t-t_0)^2/\sigma^2}.$$  

(2)

The temporal spread of this function, $\Delta t$, defined in terms of second moments, is $\sqrt{\pi/2}\sigma$, and a complementary relation is found for the frequency spread of its Fourier transform: $\Delta f = (1/\sqrt{\pi})/(1/\sigma)$. Together, they define the uncertainty principle $\Delta f\Delta t = 1/2$. For all other windowing functions, $\Delta f\Delta t > 1/2$. Throughout the text, the term bandwidth refers to $\Delta f$.

Each frequency $f$ of the Gabor transform provides one “channel” in the IFD analysis. In polar form,

$$G_w(t,f) = a_w(t,f)e^{i\phi_w(t,f)}.$$  

(3)

The instantaneous frequency of each channel is defined as

$$f'_w = \frac{1}{2\pi} \frac{\partial \phi_w(t,f)}{\partial t}.$$  

(4)

For each channel, instantaneous frequency can be estimated from the local period of oscillation, drawn from intervals between maxima or zero crossings of the signal. In this form, instantaneous frequency is calculated from information homologous to the intervals between phase-locked spikes in the auditory nerve. Instantaneous frequency is calculated analytically as follows:

$$\frac{\partial \phi_w(t,f)}{\partial t} = \frac{\partial \text{Im}(\ln(G_w(t,f)))}{\partial t} = \text{Im}\left[\frac{\partial G_w(t,f)}{G_w(t,f)}\right].$$  

(5)

From this expression, a formula in terms of the windowing function $w$ and its derivative $w'$ follows:

$$f'_w(t,f) = \text{f} - \text{Im}\left[\frac{G_w(t,f)}{G_w(t,f)}\right] \frac{1}{2\pi}.$$  

(6)

The current method is designed for signals that are tonal, defined in terms of smooth, time-dependent frequencies $F_k(t)$ and amplitudes $a_k(t)$ as follows:

$$s(t) = \sum_{k=1}^{N} a_k(t)\sin(\phi_k(t)),$$  

(7)

$$\phi_k(t) = 2\pi \int_{\tau=0}^{t} F_k(\tau)d\tau.$$  

(8)

A signal of this form is separable if the Gabor transform, at each time and frequency, receives significant energy from only one tone [one element of the sum in Eq. (7)]$^5$ Signals analyzed in this method must be separable, and must have limited frequency and amplitude modulation rates. For separable signals with sufficiently slow frequency and amplitude modulations, instantaneous frequencies $f'_w(t,f)$ of a well-chosen bandwidth provide excellent estimates of the frequency contours of the signal, $F_k(t)$. This is demonstrated in the following sections.$^5$

We use the term sparse to refer to separable signals that are modulated slowly enough to be resolved though instantaneous frequency analysis. Instantaneous frequency decomposition provides a method for finding the optimum bandwidth of analysis, and estimating $a_k(t)$ and $F_k(t)$, the amplitude and frequency contours of each component. If the method is applied to signals that are not separable, or signals with frequency and amplitude modulations that are too fast, the signal is not resolved, instantaneous frequencies do not track the signal frequencies $F_k(t)$, and the method signals its own error. The following sections illustrate what this means.

One class of test signals used in this article consist of a sum of tones with periodic frequency modulations:

$$e^{in\phi}e^{i(A/\omega)\cos(\omega t)}.$$  

(9)

Through the Jacobi–Angel expansion, a periodically modulated tone can be represented as a single frequency accompanied by an infinite sum of sidebands:

$$e^{i(A/\omega)\cos(\omega t)} = \sum_{n=-\infty}^{\infty} i^nJ_n\left(\frac{A}{\omega}\right)e^{in\omega t},$$  

(10)

where $J_n(z)$ are the Bessel functions of the first kind. This relationship is referred to in the following sections.

B. Instantaneous frequency decomposition

The method of instantaneous frequency decomposition consists of a central processing structure, an outer optimization loop, and a final quality check. The central processing structure computes instantaneous frequencies for channels of a filter bank of fixed bandwidth and applies a cross-check between tonotopic and phase information to determine which filters contribute to the analysis. The optimization loop compares frequency estimates from neighboring channels to generate a measure that we call consensus, and uses this measure to optimize the analyzing bandwidth $\Delta f$. The quality check uses the same measure of consensus to indicate specific regions of the time–frequency plane where the signal is well resolved, and other regions where high spectral density leads to a failure of the frequency estimates.

1. Raw instantaneous frequency analysis

In the first stage of the analysis, $|G_w(t,f)|$ is computed for each time and a dense set of frequencies, according to Eq. (1), for some initial choice of bandwidth. In this analysis, the distinct values of $f$ are referred to as “channels.”

Instantaneous frequency representation involves a remapping of the amplitudes $|G_w(t,f)|$ to new positions in the time–frequency plane, namely $(t,f'_w(t,f))$, where $f'_w$ is the instantaneous frequency of channel $f$ at time $t$, calculated...
from Eq. (6). This first step, the raw instantaneous frequency analysis, has been described in detail elsewhere.\textsuperscript{5}

When applied to separable signals with slow modulations, positions \((t,f)\) that are far from the signal tones are mapped onto the signal tones. This is illustrated in the following figure. Figure 1 contains an analysis of two signals according to this remapping rule. The first signal consists of two equal amplitude tones, each of which is frequency modulated with a peak to peak modulation depth of 70 Hz, over a period of 14 ms. The second signal is white noise. The frequency estimates generated from each channel provide one continuous line in each figure. [To avoid confusion, note that in panel (a), many lines overlap, leading to the appearance of a continuous distribution.] For the white noise signal, each channel responds to a slightly different portion of the white-noise spectrum, leading to a spread in frequency contours estimated from neighboring channels. The structure of this web of lines is sensitive to the bandwidth of the filter bank.

2. Tonotopic cross-check

The darkest lines in Fig. 1, panel (a), fall on the correct frequency contours of the signal. The lighter gray lines that deviate from the correct contour are generated by filters whose central frequencies are far from the primary frequencies in the signal. A qualitative explanation of this is as follows: for an unmodulated tone, off-center filters perfectly detect the true frequency, but for modulated tones, off-center filters distort the signal. Modulated signals have a broad frequency spread [Eq. (10)], and off-center filters truncate this broad frequency representation more drastically than centered filters.

The signal representation is improved by establishing a notion of “jurisdiction” for each channel. Whenever instantaneous frequency \(f^i_n(t,f)\) is far from the center of channel \((j)\), this estimate is discarded. That is, for \(|f^i_n(t,f) - f| > C\), the local estimate \(f^i_n(t,f)\) does not contribute to the analysis. The constant \(C\) we call the locking window. When this criterion is applied to a dense array of filters, discarding channels that are not “in lock” is no loss, since for each portion of the signal, there is some channel that is positioned correctly. (The number of channels locking onto a single pure tone is the redundancy of the filter bank, and in this manuscript redundancies are on the order of 10, so every frequency region is densely covered with similar filters.)

Figure 2 illustrates the effect of applying this criterion to the signals analyzed in Fig. 1. Any instantaneous frequencies coming from outside the locking window \(C = \Delta f/2 = 110 \text{ Hz}\) are not drawn in the figure. Each panel in this figure contains an equivalent number of channels, but in Fig. 2, panel (a), most channels are excluded by the locking criterion. Those that remain in the analysis condense onto two frequency contours. In contrast, the spread of frequencies in the analysis of the white noise signal [panel (b)] indicates a failure of agreement among neighboring channels, and thus a violation of the central assumption that the signal is sparse. Simple though it may be, this “blind” cross-check between tonotopic and phase information significantly improves the analysis of rapidly modulated sparse signals.

C. Bandwidth optimization through consensus

The previous section describes analysis at a fixed bandwidth. To further optimize the analysis, particularly for a signal with unknown properties, this bandwidth must be adjusted to the signal. Figure 3 illustrates the result of various bandwidth choices in the analysis of a two-tone signal. For the standard representation of the signal, the optimum filter width yields Fig. 3, panel (b). For this signal, a range of filter widths around this optimum yield the same time–frequency analysis (not shown). Much wider filter widths as in Fig. 3, panel (a), introduce interactions between the two tones, and much narrower filter widths \([c\) and \(d]\) yield a gradual transition from the modulated tone representation to the sum of sideband representation defined by Eq. (10). In panels \(a\) and \(c\), poorly matched filters lead to detailed structures of lines that are sensitive to the precise bandwidth of the analysis—a “fragile” representation of the signal.

Bandwidth is optimized by minimizing the linewidth, or consensus of the frequency estimates. This optimization can utilize a number of different objective measures of channel consensus. In this article, consensus is defined in terms of the interval between instantaneous frequency estimates from neighboring channels that are “in lock” (previous section). Specifically, consensus is the median value of \(1/[|f^i_n(t,f_a) - f^i_n(t,f_b)|\), where \(f_a\) and \(f_b\) are center frequencies for neighboring channels that are “locked” at time \(t\).
The cross-channel consensus is plotted as a function of the filter bank bandwidth for the two-tone signal analyzed in Fig. 3. The optimum bandwidth is derived from the consensus maximum. For the sparse signals analyzed in this paper the optimum bandwidth for the signal discussed in Fig. 3. For large intervals between tones and fast modulation rates, the tones overlap and error in the analysis increases. To generate the figure, the optimum bandwidth is first determined for each signal by maximizing the median instantaneous amplitude over all channels. For this reason, the magnitude of the cross-channel consensus indicates the degree of error in the analysis.

This measure performs best when frequency estimates with insignificant amplitude are excluded from the calculation of consensus. In practice, information is drawn only from channels whose instantaneous amplitude is greater than the median instantaneous amplitude over all channels.

Figure 4 demonstrates that this measure is maximized at the optimum bandwidth for the signal discussed in Fig. 3. The cross-channel consensus plotted as a function of the filter bank bandwidth for the two-tone signal analyzed in Fig. 3. The optimum bandwidth is derived from the consensus maximum. The cross-channel consensus is plotted as a function of the filter bank bandwidth, for the two-tone signal analyzed in Fig. 3.

D. Quality checks through consensus

This analysis can be applied to sparse sounds—sounds whose tonal components are separable and modulated sufficiently slowly. Fast modulations imply extended frequency representations, so modulated tones with separable center frequencies may nevertheless have significant frequency overlap due to their modulations. To illustrate why fast modulations require wideband analysis, consider a pure tone at frequency \( \omega \) that is periodically modulated in amplitude at a lower frequency \( \omega_2 \). This signal, \( \cos(\omega t)\cos(\omega_2 t) \), is equivalent to \( \cos(\frac{1}{2}\omega t + \frac{1}{2}\omega_2 t) + \cos(\frac{1}{2}\omega t - \frac{1}{2}\omega_2 t) \), and to accurately represent it within a single band, a filter centered at \( \omega \) must have a frequency bandwidth of at least \( 2\omega_2 \).

Similarly, Eq. (10) reveals that a single tone with periodic frequency modulation involves a sum of sidebands with an infinite extent in frequency. Any bandpass filtering will involve truncations of the sum, and the severity of the truncation depends on the center frequency and bandwidth of the filter, as well as the time scale and amplitude of the modulations. If a signal is sparse by our definition, the truncation of frequency modulations at the optimum bandwidth is negligible, and neighboring channels produce very similar results. Alternatively, if the signal is not sparse, truncation is significant, leading to distinct frequency estimates in different channels. For this reason, the magnitude of the cross-channel consensus indicates the degree of error in the analysis.

Figure 5 demonstrates a correlation between cross-channel consensus and frequency error for a family of test signals. Each signal in the set consists of two frequency-modulated tones separated by a fixed interval, as illustrated in Fig. 3. For large intervals between tones and slow modulation rates, the signal is spectrally sparse and can be resolved with the IFD method. For small intervals between tones and fast modulation rates, the tones overlap and error in the analysis increases. To generate the figure, the optimum bandwidth is first determined for each signal by maximizing consensus, as described in the previous section. At the optimum bandwidth, rms error between the known signal content and the IFD estimate is plotted against the median consensus value over the time–frequency plane. When modulation rates are too fast to resolve, consensus measures decrease.

In addition to averaged quality measures, consensus within local regions of the time–frequency plane can indicate well-resolved signal components within a larger analysis. (Even the white noise analysis in Fig. 2 displays what appear to be “caustics,” or regions of high agreement between nearby filters, though the overall analysis is characterized by low consensus.)

In summary, consensus between redundant channels is used to guide bandwidth optimization, and to signal the quality of the final analysis. In principle, local consensus measures can be used to find spectrally sparse components within more complex signals, or to adapt a bandwidth separately for different regions of the time–frequency plane.
III. RESOLUTION AND PRECISION

Understanding the limits of the method requires introducing a distinction between resolution and precision, a distinction well developed in, e.g., microscopy. Precision is the accuracy with which the position of a given object can be computed, while resolution is the smallest distance at which two objects may be discriminated as distinct.

A. Resolution

The IFD analysis requires that the bandwidth of the filters be narrower than the separation between adjacent frequency components. Since the time accuracy of the analysis is inversely proportional to bandwidth, the IFD resolution is constrained by a variant of the Fourier uncertainty relation:

$$\Delta T \Delta f_{\text{min}} \geq \frac{1}{2},$$

where \(\Delta f_{\text{min}}\) is now the minimum separation between adjacent frequency components, and \(\Delta T\) the effective time resolution with which frequency changes can be tracked.

The resolution limit of the IFD method can also be described in terms of the maximum modulation rates that can be resolved for a given separation of frequency components. One example of this limit is as follows: for frequency modulations that are faster than the depth of modulation \((A \leq \omega)\) in Eq. (10), \(\Delta f/2\) must be greater than the modulation rate \(\omega\), otherwise sidebands of the modulated signal are severely truncated.

B. Precision

As for other methods specialized for sparse sounds, IFD can achieve high precision in both time and frequency whereas general Fourier analysis is limited by the uncertainty principle. For example, frequency errors for the signals analyzed in Fig. 5 range from less than 1 Hz to 10 Hz, whereas the time scale of modulations in these signals imply a classical frequency uncertainty as high as 300 Hz.

Figure 6 contains an explicit comparison with classical frequency uncertainty for a family of test signals. To produce

FIG. 6. A comparison of general time–frequency precision based on spectral derivative pitch tracking (SD) with IFD estimates. The rms error in pitch tracking is plotted as a function of the modulation rate of the test signals. The analyzed signal contains modulated tones centered on 1100, 2000, and 2900 Hz. Each tone is independently modulated with fast frequency modulations—200 Hz peak to peak in panel (a) and 40 Hz peak to peak modulations in (b). The optimum time scale of the windowing function in the Fourier analysis was determined and used in this comparison (21 ms). (In the spectral derivative analysis, the windowing functions are prolate spheroidal sequences.) The fixed bandwidth IFD analysis uses Gaussian windows of duration 1.6 ms. The IFD analysis (like many other methods adapted to sparse signals) achieves a pitch tracking precision that can be orders of magnitude sharper than the resolution of general Fourier analysis.

FIG. 7. Vocal illustrations: analysis of a whistle in a canary song. Panel (a) contains a windowed short-time Fourier analysis or sonogram with the following parameters: analyzing window 23 ms, 80% overlap. Panel (b) contains an IFD analysis with channels of bandwidth \(\Delta f/2 = 600\) Hz, spaced 20 Hz apart. The locking window for this analysis is \(\Delta f/2\). Panels (c) and (d) contain close-up views of a frequency instability in the whistle. Pixel intensities for all four panels were scaled from white (30 dB below the maximum power) to black (maximum power) according to the logarithm of signal power. The fundamental resolution of classical time–frequency analysis \((\Delta f/\Delta > 1/2)\) is indicated by the gray rectangle in panel (c).
this figure, frequency contours of the test signals are estimated based on either IFD or a short-time Fourier method (zero crossings of multitaper spectral derivatives\(^{25-27}\)). The rms error in frequency contour estimation was then calculated. Over a range of modulation rates, the IFD analysis at a fixed bandwidth achieves a precision of frequency estimation one or two orders of magnitude sharper than the resolution of general Fourier analysis.

Enhanced resolution for sparse signals is not surprising. Any method specialized for sparse sounds will outperform a more general time–frequency analysis. For specific signal ensembles, specialized applications of Fourier analysis can also outperform the limits of the general method. For example, in the analysis of sparse signals, frequency contours can be more precisely localized by interpolating the Fourier estimates between frequency bins.\(^{28}\) Comparisons have been made among methods of Fourier interpolation\(^ {29,7}\) and measures of instantaneous frequency. In the vicinity of a spectral peak, instantaneous frequency measures meet or exceed the precision of pitch tracking achieved through Fourier interpolation.\(^ {29,7}\)

Relative to other specialized methods, the primary advantage of the IFD method is the generality conferred by redundancy and cross-check. No information is needed about the analyzed signal to apply the method. If the signal is sufficiently sparse, an optimized analysis is found without reference to the signal character.

### IV. CONCLUSION

IFD represents sparse signals in time and frequency with high precision through a self-optimized instantaneous frequency analysis. Two aspects of cross-validation are employed to optimize the analysis. The tonotopic cross-check compares tonotopic and phase information within each channel. A filter contributes locally to the analysis only if its center frequency and instantaneous frequency match. In a second cross-check, the consensus of frequency estimates from neighboring channels is used to guide the optimization of analysis bandwidth for a given signal, and to signal the degree of error in the analysis. When applied to sparse signals, the redundant channels of the IFD generate high-con-
nels with overlapping passbands generate similar spike patterns, and the end stage, but at the computational cost of highly redundant channels fail to coincide at any bandwidth, signaling the breakdown of the method.

The elements of this analysis may be relevant to auditory processing. Many animal vocalizations contain well-defined pitches that are rapidly modulated, and neural auditory processing has evolved under a need to make demanding distinctions in both time and frequency simultaneously. To achieve an optimum representation of sparse sounds, IFD provides a rationale for integrating information from tonotopic and phase information in the auditory nerve. Cross-checks between spike intervals and the tonotopic position of a fiber could select the fibers with optimal center frequencies. A similar criterion was employed by Srulovicz and Goldstein to explain psychophysical data for the perception of simple unmodulated signals. Second, confidence can be placed on a frequency estimate when different channels with overlapping passbands generate similar spike intervals. Among a redundant set of nerve fibers with varying bandwidths, the cells that form a consensus in their interspike intervals may stand out as salient, preferentially drawing information from channels whose bandwidth was well suited to the local signal content. As early as the cochlear nucleus, there are cells that receive inputs from auditory fibers with a range of center frequencies, thus at this stage of auditory processing or beyond, measures of cross-channel consensus could in principle be implemented.

The method of cross-channel comparison has the potential for high compression and top reconstruction quality at the end stage, but at the computational cost of highly redundant arrays of sensors, and a large number of cross-channel comparisons. Early expansive stages in the neural pathways of hearing and vision may serve a similar function: to provide higher accuracy and efficiency not at intermediate stages, but at the far end of the processing pipeline.

**ACKNOWLEDGMENTS**

The authors would like to thank A. Libchaber, F. Nottebohm, A. J. Hudspeth, and P. P. Mitra for their comments on the manuscript. This work was supported in part by the Burroughs Wellcome Fund.