Time-Frequency Analysis of Musical Signals

WILLIAM J. PIELEMEIER, MEMBER, IEEE, GREGORY H. WAKEFIELD, MEMBER, IEEE, AND MARY H. SIMONI

Invited Paper

The major time and frequency analysis methods that have been applied to music processing are traced and application areas described. Techniques are examined in the context of Cohen's class, facilitating comparison and the design of new approaches. A trumpet example illustrates most techniques. The impact of different analysis methods on pitch and timbre examination is shown. Analyses spanning Fourier series and transform, pitch synchronous analysis, heterodyne filter, short-time Fourier transform (STFT), phase vocoder, constant-Q and wavelet transforms, the Wigner distribution, and the modal distribution are all covered. The limitations of windowing methods and their reliance on steady-state assumptions and infinite duration sinusoids to define frequency and amplitude are detailed. The Wigner distribution, in contrast, uses the analytic signal to define instantaneous frequency and power parameters. The modal distribution is shown to be a linear transformation of the Wigner distribution optimized for estimating those parameters for a musical signal model. Application areas consider analysis, resynthesis, transcription, and visualization. The more stringent requirements for time-frequency (TF) distributions in these applications are compared with the weaker requirements found in speech analysis and highlight the need for further theoretical research.

I. INTRODUCTION

A physical example, such as a whistle, is useful when first learning the concept of frequency. The teacher purses his lips, blows once, changes the position of his tongue, and blows again. One hears two sounds, a low-pitched followed by higher-pitched whistle, and is told that the power spectra of these two signals differ in that the first sound concentrates the distribution of its acoustic power in the neighborhood of a frequency substantially lower than that of the second. Thus the mathematical expression $f_1 < f_2$ is learned through our everyday experience of pitch, i.e., the lower the frequency, the lower the pitch.

The lesson may be good engineering pedagogy but it is definitely very poor music processing: actual power spectra of the whistles reveal that both signals are wideband and that the power in each whistle sample $w_i(t)$ is concentrated in the neighborhoods of an *ordered set* of isolated frequencies, $\{f_{i0}, f_{i1}, \cdots, f_{iM}\}$, for $i = 1, 2$, where $f_{ij} < f_{i,j+1}$ for all i and j. Musical acoustics refers to this set as the instrument's *partials* and in certain cases, such as a good "whistleblower," the partials are *harmonically* related to a *fundamental* as in the relationship $f_{ij} = jf_{i0}$ for all $j > 0$. Thus the partials in a harmonic set are integer multiples of the fundamental, f_{i0}.

The above correspondence between frequency and pitch is identical only for monochromatic signals, e.g., sinusoids, such as those generated by an oscillator for which the set of partials consists entirely of one frequency, $\{f_0\}$. When comparing the pitch of instruments with many harmonics, their fundamentals, rather than their wideband set of partials, determines the pitch so that, in general, the lower the fundamental, the lower the pitch. That we are sensitive to the harmonic relationships among partials is reinforced by the phenomenon known as the "missing fundamental"; even when the fundamental is absent from the set of harmonics, we hear it as if it were present [1].

When discussing musical signal processing, we must always be aware of the differences between the mathematics of representing an acoustic signal and our auditory perception of that signal. To ignore these differences discounts the interesting phenomena of our auditory perception. What, after all, is more mathematically elegant than to represent the pitch of a whistle using the basic unit of our Fourier mathematics, frequency! But how interesting to discover that it is not the frequencies of the partials *per se*, but their relationships, that we perceive as pitch; only when these are lacking, as in the case of a single sinusoid, is our Fourier mathematics adequate to represent the pitch that we hear.

Manuscript received July 24, 1995; revised February 29, 1996. G. H. Wakefield's work is supported in part by grants from NIH, ONR, the Ford Motor Company, and the Office of the President of the University of Michigan.

W. J. Pielemeier was with the Department of Electrical Engineering and Computer Science, The University of Michigan, Ann Arbor, MI 48109. He is now with the Ford Motor Company, Scientific Research Laboratory, Dearborn, MI 48121 USA (e-mail: w.pielemeier@ieee.org).

G. H. Wakefield is with the Department of Electrical Engineering and Computer Science, College of Engineering, The University of Michigan, Ann Arbor, MI 48109 USA (e-mail: ghw@eecs.umich.edu).

M. H. Simoni is with the Center for Performing Arts and Technology, School of Music, The University of Michigan, Ann Arbor, MI 48109 USA (e-mail: msimoni@umich.edu).

Publisher Item Identifier S 0018-9219(96)06202-0.

The inadequacy of Fourier mathematics to succinctly represent perceptual phenomena of musical signals motivates the theory of time-frequency (TF) representations. The whistle provides a case in point. Rather than two distinctly pitched events, suppose the instructor had slowly varied the volume of the vocal cavity from large to small while blowing, thereby creating a whistle whose pitch varied over time. A Fourier representation of this event is inherently flawed, lacking the element of time, in contrast with the first case where the representation preserves the relationships among the partials which form the whistle's unchanging pitch. Indeed, a power spectrum of the temporally varying pitch reveals that the well-known utility of the Fourier mapping of functions of time into ones of frequency is precisely *not* what we want when representing signals which change in their pitch over time, since what we hear suggests a representation in both time and frequency. Ignoring pitch for the moment, a time-variant whistle beckons a three-dimensional (3-D) plot that represents frequency, amplitude, and time. As we shall argue in Section II, the discipline of TF representation is a mathematical response to the need for such mappings in representing our auditory response to acoustic signals.

The use and advancement of such a theory in musical signal processing reflects our interests in quantifying the properties of those signals we call music so that we may better understand the processes involved in the generation and appreciation of music. Such interplay between mathematics, engineering, physics, psychology, and aesthetics is not new; the history of music signal processing dates well into antiquity. Indeed, many mathematical physicists have created their systems, in part, in response to their fascination with music. From the fundamental relationships among length, vibration, and pitch established by Pythagoras to the development of partial differential equations in the 19th century by Helmholtz, Rayleigh *et al.*, music has provided a wealth of inspiration to what we now consider to be standard engineering mathematics and physics. It is our belief that music has helped carve out our current understanding of and need for TF distributions, as it has for centuries. Through a confluence of technological, theoretical, pedagogical, and economic factors, over the next decade researchers will be able to add new layers of mathematics to the substrates that music provides.

The primary scope of the present paper is limited to a discussion of TF distributions in music processing. We argue that current work in this area is best understood when presented in the historical context of the research problem. Within this development, Section II establishes a common framework for various analysis techniques, and covers one-dimensional (1-D) methods which predate the use of joint time and frequency variables. Section III continues with the application of joint TF analysis to musical signals. Throughout this historically oriented development, we will use one particular acoustic source as the example by which we illustrate the evolution of our ideas of musical instruments from the "steady-state" notions of Helmholtz to the "dynamic time-varying" constructs of the past two decades. While no one instrument can best exemplify all mathematical models, the trumpet sample we have selected possesses many musical source features that reveal the pitfalls of various mathematical representations.

In Section IV, we widen the scope of the presentation to discuss the variety of applications, building upon analysis, that presently drive music processing. There are numerous parallels between this and speech processing. We summarize these correspondences at a relatively high level.

II. Time-Frequency Distributions: Historical Precursors in Music Processing

We introduce a common mathematical framework for representing the fine temporal and spectral structure of musical signals. This enables us to directly compare representations starting with that of Helmholtz, and ranging through common Fourier techniques, to more recent techniques such as wavelets and bilinear transforms. In this framework we see the historic problem faced by all these approaches to music analysis concerning pitch and timbre, which debates the dual spectral and temporal views of the signal. We consider a trumpet sample from the McGill University Master Samples of individual notes on compact disc [2] and view it in the context of the various distributions we cover. This sample contains the attack and the initial steady state of a note with a pitch of 185 Hz.[1] It illustrates spectral components with rapidly varying frequencies and amplitudes which highlight the ability of various representations to capture such information.

A. Generic Mathematical Form

The generic form that we propose for comparing musical signal representations is a joint distribution of signal energy over time and frequency. The term *distribution*, in this context, is applied to purely deterministic signals and refers to an intensity per unit time and frequency, rather than any probabilistic interpretation. A musical signal is given as a function of time, with frequency and energy variables chosen for the useful information that they provide in our context, and because most of the distributions that we consider can be written in these terms. Many of the distributions, such as the Fourier series (FS) and Fourier transforms (FT), and constant-Q transforms, have coefficients which are converted to units of energy by computing the squared magnitude. Others, such as the Wigner distribution (WD) and modal distribution (MD), possess coefficients which already have units of energy. Thus an energy function provides a common basis for comparison. It is also a natural variable for analyzing the physics and psychophysics of musical sources, and is useful in resynthesis.

The majority of the two-dimensional (2-D) systems for depicting musical signals that we consider have frequency

[1] The note chosen is six semitones below middle C, where middle C has a pitch of 262 Hz. which is sometimes designated F#3. It is taken from Volume Two, Track 16, Index One of the McGill University Master Samples.

as an independent variable. Those that do not can typically be recast as functions of frequency or given a frequency interpretation. The wavelet transform, a function of time and scale, provides an example of this. It has been shown [3], [4] that a wavelet transform has a corresponding frequency interpretation, with the frequency varying inversely to the scale. Thus a frequency variable encompasses all of the analysis methods that we consider, and combined with time and energy, defines a general framework in which to compare and design suitable representations for musical signals.

The work of Wigner [5] in the probabilistic context of quantum mechanics laid down the framework upon which TF representations were later refined. Ville [6] subsequently derived a deterministic TF distribution which was formally analogous to Wigner's. The Wigner–Ville distribution of a signal $s(t)$ is given by

$$W_s(t,\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} s\left(t+\frac{\tau}{2}\right)s^*\left(t-\frac{\tau}{2}\right)\exp(-j\omega\tau)\,d\tau. \tag{1}$$

Early fundamental work on TF distributions was also done by Gabor [7], Page [8], *et al.* In his historical review of the subject, Cohen [9] says:

> Although it is now fashionable to say that the motivation for this approach is to improve upon the spectrogram, it is historically clear that the main motivation was for a fundamental analysis and a clarification of the physical and mathematical ideas needed to understand what a time-varying spectrum is.

These pioneers' work was further built upon by Cohen, who realized that an infinite class of TF distributions, including those of interest to us, could be obtained from the Wigner–Ville distribution by linear transformation [10]. One expression of this linear relationship is

$$C_s(t,\omega;\varphi) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} W_s(\tau,\xi)\varphi(t-\tau,\omega-\xi;t,\omega)\,d\tau\,d\xi \tag{2}$$

where $C(t,\omega;\varphi)$ is a member of Cohen's class, and $\varphi(\tau,\xi;t,\omega)$ is called the kernel of the distribution [11]. Cohen showed that the kernel, an arbitrary function which defines the particular linear transformation in (2), is simply related to many important properties of the distribution, and thus can be used to test or constrain those properties. We choose Cohen's class as our framework because it forms a basis for both the quantitative analysis of existing distributions in musical signal processing and the design of new ones with desirable properties.

Alternatives classes such as the affine class of energy functions of time and scale [4] provide a less appropriate context for our work. A frequency variable is more directly related to auditory pitch and more commonly used in musical acoustics and synthesis than the scale variable, and these are important potential applications for the results of musical signal analysis. Furthermore, most of the commonly used musical signal analysis techniques are included in Cohen's class, but not the affine class. The constant-Q and windowed exponential wavelet transforms, among the few musical signal analysis techniques which do exhibit scale properties, may also be interpreted as functions of frequency [3], [4]. These transforms can also be cast in Cohen's class by exploiting the often-ignored contingency for frequency dependent kernels in Cohen's class [9], [11], and identifying scale with the inverse of frequency. Thus Cohen's class is clearly the more inclusive choice for our work.

B. Helmholtz and the Fourier Series

Helmholtz [12] built upon the concepts of Fourier, Bunsen, and Ohm. While it was Fourier who first developed the theory of complete orthonormal expansions of signals as sums of sinusoids, the driving force behind this development had little to do with what we now call Fourier analysis and much more to do with the operational properties of the transform as the tool for solving certain boundary value problems in partial differential equations [13]. It was Bunsen who saw in the mathematics of the Fourier theory a means for decomposing signals into sums of "simpler signals," and Ohm who, in his acoustic law, proposed that the auditory system extracted the amplitudes of these same simple signals when perceiving acoustic signals [1].[2]

Though Helmholtz knew that musical sounds varied over time, generally containing attacks and decays, he believed that the salient portion from a perceptual standpoint was nearly periodic, and called it the steady state. The FS theory states that repeating or periodic functions can be broken down into a sum of sinusoidal components [14]. The repetition rate defines a fundamental frequency, with all of the sinusoidal component frequencies, called harmonics, at integral multiples of the fundamental frequency.

Helmholtz believed that the magnitudes of these harmonic components determined loudness, pitch, and timbre, which he considered to be the primary perceptual attributes of individual musical sounds. The pitch was defined as the single sinusoidal frequency which a complex tone is perceived as most similar to, while timbre accounted for all differences in perception outside of loudness and pitch. Though loudness is now fairly well understood, even today the investigation of pitch and especially timbre is far from a closed subject, and these are still considered primary perceptual attributes of musical tones. Helmholtz believed that pitch was determined by the fundamental frequency of the harmonic series, and timbre was determined by the pattern of magnitudes of the remaining components. His limited spectral analysis tools, including acoustical cavity (Helmholtz) resonators, tuning forks, and reeds, provided some information about the amplitude and frequency of harmonic components, but no information about the time varying nature of the sound.

While Helmholtz's analysis techniques estimated component amplitudes and frequencies directly, the closest

[2] Ohm's acoustic law is not the same as the more famous of Ohm's laws relating electrical quantities.

mathematical technique is the FS, which he relied upon as the theoretical basis for his analysis. The FS coefficients for the periodic signal $s(t)$ with fundamental frequency ω_0 are computed with the following integral over one period of the fundamental frequency

$$S(n) = \frac{\omega_0}{2\pi} \int_{-\pi/\omega_0}^{\pi/\omega_0} s(t) \exp(-jn\omega_0 t)\, dt \qquad (3)$$

where n refers to the harmonic number [14]. Thus the FS is a function of a discrete set of harmonic frequencies only, computed based on a single period of the signal, and is not a function of time. In the case of musical signals which are not strictly periodic, an average or estimated pseudo-period is sometimes used in the computation [15][16], but then the conclusion that the signal contains strictly harmonic components is no longer assured, and (3) becomes more difficult to interpret. The energy at each harmonic is obtained by computing the magnitude squared $|S(n)|^2$, or power spectral density, of the FS coefficients.

To place the FS power spectral density in the Cohen's class framework, we note that it is a function of frequency only, and therefore comparable to the frequency marginal $P_s(\omega)$. The frequency marginal is obtained from a distribution by integrating out the time variable [9], leaving a 1-D distribution given by

$$P_s(\omega) = \int_{-\infty}^{\infty} C_s(t, \omega; \varphi)\, dt. \qquad (4)$$

The frequency marginal in (4) is sampled at the discrete frequencies $n\omega_0$ to compare it to the power spectral density $|S(n)|^2$. Fig. 1(a) illustrates the power spectral density of a FS analysis from sampled data, showing the first 20 harmonics of the trumpet note described at the outset of this section, with pitch and approximate fundamental frequency of 185 Hz. The 5.4 ms (1/185) segment starts 500 ms following note onset, where the steady state portion begins. The individual periods of trumpet samples are generally very impulsive, with the result that the sample contains many harmonics, extending beyond the plot in Fig. 1(a) to 50 harmonics in many cases.

If the frequency marginal of a distribution matches the power spectral density, the distribution is said to satisfy the frequency marginal. All else being equal, this is a desirable property for a distribution, because it means that it is consistent with that power density. In practice, however, this property trades off against other desirable properties, and most of the distributions in common use for musical signal analysis do not satisfy the frequency marginal. The WD does satisfy its frequency marginal [17], lending to its special status in Cohen's class.

The Helmholtz analysis raises the question of whether spectral analysis based on a single period fully accounts for the perceptual attributes of even individual musical notes. It is now understood that the attack portion of a musical tone is important to correct instrument identification [18]. Both pitch and timbre are known to depend upon joint temporal and spectral characteristics of the signal [18], [19]. Furthermore, strict periodicity generally does not hold even during the steady state, and thus the component magnitudes computed over a single period give at best an incomplete representation. Additionally, during segments of rapid change such as attacks, the FS gives inaccurate estimates of magnitude and frequency even for a single time [20]. Thus while Helmholtz's research accomplished a great deal given the available techniques, it only opened the door to a much richer set of questions concerning the joint role of time with frequency.

C. Long-Term Spectrum Analysis and Wideband Envelope Analysis

The spectral composition of the entire signal $s(t)$, rather than just one period, can be obtained by the FT [14]. This is computed by

$$S(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} s(t) \exp(-j\omega t)\, dt. \qquad (5)$$

Here the power spectral density is the magnitude squared $|S(\omega)|^2$ over the continuous frequency variable ω. While this provides information over more frequencies and about a longer time interval than the FS, it still lacks temporal information about any additional spectral content. The power spectral density of the FT is, like the FS case, most directly comparable to the frequency marginal for Cohen's class computed in (4). Fig. 1(b) shows part of the FT power spectral density of our trumpet sample, displaying the frequency region of the 16th–19th harmonics, which may be compared to the 20 harmonics in Fig. 1(a). In the FT case, the entire 500 ms of our sample is analyzed, rather than just a single 5.4 ms period as in Fig. 1(a), with the inclusion of the attack and other nonstationarities contributing to the additional spectral content between harmonic frequencies which appears in Fig. 1(b).

We have pointed out that the FS and transform lack information about the distribution of signal energy over time. A 1-D view of temporal variation is provided by the instantaneous power $|s(t)|^2$. Instantaneous power corresponds most directly to the 1-D time marginal of a 2-D Cohen's class distribution, obtained by integrating out the frequency variable in

$$p_s(t) = \int_{-\infty}^{\infty} C_s(t, \omega; \varphi)\, d\omega. \qquad (6)$$

Paralleling the case with the frequency marginal, we say that a distribution satisfies its time marginal if that function is equal to the instantaneous power. Most distributions used for musical signal analysis do not satisfy their time marginals. Again, the WD does, providing another reason for its central role in Cohen's class. Just as the power spectral density gave no information about how the signal energy is distributed over time, the instantaneous power provides none over frequency.

The instantaneous power of our trumpet sample is displayed in Fig. 1(c), calculated over 17 periods from 75 to 165 ms into the sample. This plot reveals that the sample is not strictly periodic, even during the latter steady state portion. The FS spectrum in Fig. 1(a) was taken over

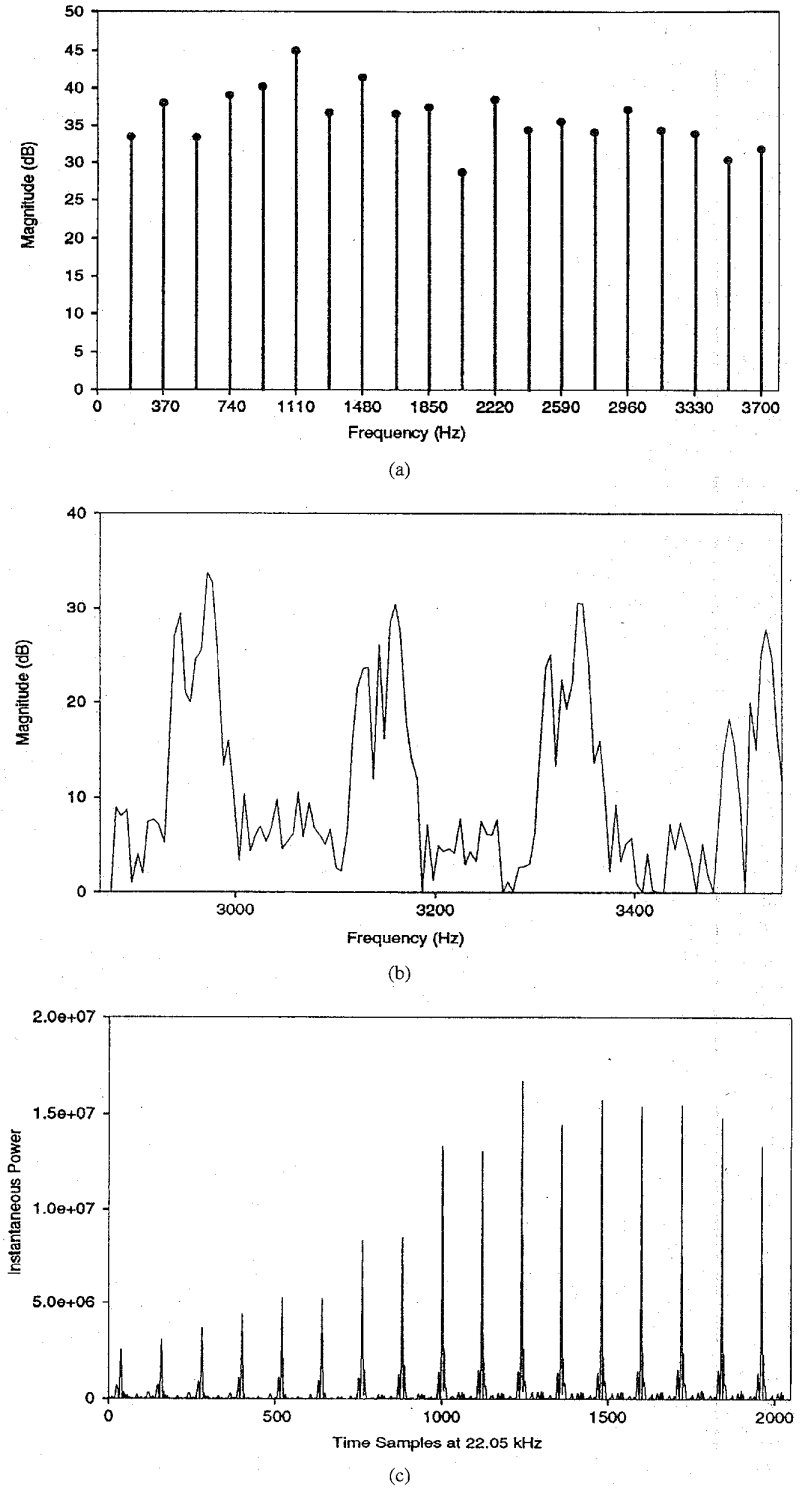


Fig. 1. One-dimensional, time, or frequency-only views of a trumpet signal: (a) FS analysis of one period, showing the first 20 components, (b) FT of the entire attack, showing the 16th–19th components, and (c) instantaneous power of the attack over $\approx$ 100 ms, encompassing 17 periods.

one period late in this sample, while the FT in Fig. 1(b) was taken over the whole sample. An alternative approach to analyzing the power trend uses the envelope of the sample, which encloses the instantaneous power peaks. The envelope of a signal or of an individual component in that signal is important in the context of time-varying signals, and is elaborated later.

A number of perceptual attributes of music are related to instantaneous power, such as amplitude vibrato, and the slope or shape of the attack and decay of a note. Even

though the instantaneous power lacks spectral information, a number of theories of pitch perception are based on strictly temporal information [19], [21]. These measure the period between similar peaks in the time signal, and help account for ambiguities in the pitch of complex tones where several competing pitch percepts may sometimes be heard.

III. Application of TF Methods to Musical Acoustics

The analysis perspectives described in Section II give insight into the distribution of musical signal energy with respect to time or frequency individually, but a joint function of these two variables is needed to provide a complete picture. A number of attempts were made in the 1950's and 1960's using analog techniques [22]–[26], but these had inherent limitations in accuracy and mathematical flexibility. The digital electronic era greatly reduced these obstacles, leading to two lines of research to construct such distributions. One focused mainly on applications, and adapted existing frequency analysis techniques for a single time interval to provide temporal information. Concurrently, the work of Wigner, Cohen, and others previously mentioned concentrated on the theoretical issues involved in constructing such joint functions. In this section we review the former techniques first, describing how these TF analyses were and are applied to musical signals. We then look briefly at a Cohen's class approach which uses the estimation of the parameters of a musical signal model as its focus, combining the flexibility of this general framework with the particular requirements of the musical signal analysis application.

A. Time-Varying FS Extensions

The first line of research that we review takes the static spectral analysis of FS and transform techniques, and adapts them to provide time-varying information; an approach with inherent limitations. FS and transforms analyze signals by decomposing them into a sum of sinusoids each having constant frequency and amplitude. Time-varying adaptations assume that the frequencies and amplitudes of the signal are nearly constant over limited time intervals, and then spectrally analyze successive intervals as if each is stationary. The assumption of stationarity over an interval is in fundamental conflict with the concepts of instantaneous frequency and amplitude needed for time-varying signals.

Techniques which extended static frequency analysis methods were often first developed for the analysis of speech signals, and later applied to music, which has related characteristics such as quasiharmonic structure and pitch. This included time-varying extensions of FS techniques. Mathews and his colleagues [27] developed the technique of pitch synchronous analysis of speech, which applied a running FS analysis whose period changed based on an estimate of the changing fundamental frequency. Luce [15] was among the first to apply such techniques to musical tones.

Luce analyzed 14 orchestral instruments using a running FS analysis similar to Mathews'. In Luce's approach, proposed by M. Clark, musical notes are digitized and then low-pass filtered to isolate the fundamental frequency, whose changing value is estimated by counting zero crossings. A FS analysis is then performed on successive pseudo-periods, whose lengths are based on those estimates. The FS in (3) is applied for fundamental periods spanning successive times t, and the running phases computed using

$$\phi_t(n) = \operatorname{atan}\left(\frac{\Im[S_t(n)]}{\Re[S_t(n)]}\right) \tag{7}$$

where $\Re$ and $\Im$ are the real and imaginary parts of the complex coefficients $S_t(n)$. Since the signal is generally not perfectly periodic, the harmonic frequencies vary from integral multiples of the fundamental, and their value is estimated from the coefficients by looking at the change in phase between periods t_0 apart, rather than by assuming them to be multiples of the fundamental as a FS analysis normally does, as given by

$$\overline{\omega}_t(n) = \frac{\phi_{t-t_0}(n) - \phi_t(n)}{t_0}. \tag{8}$$

The magnitudes of the harmonics are computed at each time by $|S_t(n)|^2$. Those magnitudes do not take into account the improved frequency estimates in (8), but correspond to the strictly integer multiple harmonic frequencies $n\omega_0$, resulting in errors.

Luce and Clark [28] applied Luce's analysis data to the additive resynthesis of brass instruments. Luce's analysis technique had flaws, including sampling the high frequency content of the signal too slowly, and this affected the quality of the resynthesis. Strong and Clark [29] also performed additive resynthesis, utilizing most of the brass and woodwind data collected by Luce. These studies and those that follow by Mathews *et al.* involved subjective testing which confirmed the importance of the time-varying signal content, and particularly the attack, in instrument recognition and timbre.

Luce and his coworkers noted a phenomena in analyzing brass instrument attacks which they called "blips," corresponding to brief amplitude peaks early in the attack. Resynthesis showed that the blips were important to a realistic brass sound, but the authors were unable to explain why the blips occurred or whether they were artifacts of the analysis. We will see that this is due to the limitations of their methods. Strong and Clark [29] comment on the difficulty in interpreting a periodic representation of a signal which is not strictly periodic.

Mathews and his colleagues subsequently followed up their earlier speech analysis efforts with pitch synchronous analyses of musical tones, avoiding many of Luce's problems. Risset and Mathews [30] used this technique to analyze and resynthesize the trumpet. Their data showed the same transient rises in components during the attack as Luce *et al.*, though they did not refer to them as blips. They also observed some shifting in frequency during the attack, and a delay in the onset of higher frequency

components, showing a limited ability to track frequency variations, but did not note any relationship between the frequency and amplitude shifts. Mathews and Kohut [31] used pitch synchronous analysis to study the violin. Their study indicated that a complex time-varying interaction between the string frequency and closely spaced violin body resonances contributed to the timbre of the instrument, another example of the role of joint TF analysis in timbre research.

The pitch synchronous analysis technique just described has two primary weaknesses for time-varying signals. When (3) computes the coefficient of a particular harmonic n for the perfectly periodic signals required of a true FS analysis, all other harmonics are perfectly canceled. However, if the signal is not perfectly periodic, other components are not perfectly cancelled [20]. Furthermore, in the ideal harmonic component case, each component is multiplied by a gain of one in the analysis and its amplitude estimated correctly. However, if it is not exactly harmonic, then it is multiplied by a gain less than one and underestimated [20]. Signals which vary rapidly in amplitude and frequency, common during an attack, are subject to these types of errors.

Beauchamp developed an extension of pitch synchronous analysis which partially dealt with the leakage problem cited above. Equation (3) implicitly corresponds to multiplying the signal $s(t)$ by a rectangular window of length $t_0 = 2\pi/\omega_0$, which is one period of the estimated fundamental frequency, and then computing a FT at only harmonic frequencies. Beauchamp [20] proposed changing this to a Hanning [14] window which is $2t_0 = 4\pi/\omega_0$ long (i.e., two periods). This window more effectively suppresses leakage from other components when they are not at exact harmonic multiples. However, the problem of gain distortion when the component is not an exact harmonic is slightly worse with the Hanning window. There is also a sacrifice in time resolution relative to pitch synchronous analysis with a rectangular window, because the parameter values are estimated based on an average over a longer time. The Hanning window is also not effective against leakage from components which are less than a harmonic multiple apart or which shift in frequency so rapidly that a meaningful estimate of the period is not possible. Beauchamp applied a variation of the Hanning window extension of pitch synchronous analysis to a major study of violin tones, further supporting Mathews and Kohut's conclusion that string vibrato interacted with a sharply peaked violin body filter characteristic to produce the violin's timbre [32].

The heterodyne filter is a variation on Beauchamp's method, and uses a single period rectangular window as with pitch synchronous analysis, but adds post filtering to the output to give an overall result similar to Beauchamp's. Moorer [33] described the heterodyne filter in his thesis work on automated music transcription, and subsequently collaborated with Grey in a study of musical timbre using it [34]. Their study confirmed the important role of attack transients, including frequency and amplitude changes of individual components, in the perception of timbre. In a related paper with Strawn, they specifically note the presence of Luce's "blips" in the trumpet attacks they analyzed [35], and note that the blips often occur as the frequency of a component stabilizes.

B. Time-Varying FT Extensions

Another adaptation of a static spectral analysis technique which originated with speech work and was later applied to music is the short-time Fourier transform (STFT) and the related phase vocoder. Paralleling extensions for FS to time-varying signals, the STFT is a time-varying extension of the FT. The STFT applies a sliding window $h(t)$ to the FT in (5) to give

$$S(t,\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} s(\xi)h(t-\xi)\exp(-j\omega\xi)\,d\xi. \qquad (9)$$

The sliding window isolates a short segment of the signal for Fourier analysis, thus providing frequency information as a function of time.

The phase vocoder converts the STFT coefficients to a magnitude function $|S(t,\omega)|^2$ and a phase function analogous to (7) for the FS. The phase vocoder was first developed by Flanagan and Golden [36] at Bell Labs for coding speech signals. Shafer and Rabiner [37] studied the STFT for the analysis and synthesis of speech. A number of important improvements to the phase vocoder and STFT involving theory, computational efficiency, and applications, were developed by Portnoff [38]–[40]. Dolson [41] developed a version which tracked the fundamental frequency of a musical tone. He also analyzed the relationship of the parameters computed by the vocoder with the instantaneous magnitudes and frequencies of the analytic signal. This is an important comparison because the analytic signal leads to a well-defined concept of time-varying frequency and amplitude. Strawn [42] used a version of the phase vocoder which predated Dolson to analyze musical transitions between notes, thus revealing another time-variant spectral characteristic which contributes to timbre perception.

The extended FS techniques described earlier are technically a subset of the STFT and phase vocoder methods where the window length is restricted to a multiple of the signal's fundamental period (usually one or two periods) and coefficients are only computed at multiples of the fundamental frequency. The extended FS windows generally have short time spans relative to the requirements of musical signal analysis. For example, middle C has a period of about 4 ms, near the minimum threshold of about 1 or 2 ms below which humans cannot hear temporal variations in components [43], and therefore analysis with a single period window provides excellent perceptual time resolution. A note with a pitch two octaves below middle C (period of about 15 ms) still affords relatively good temporal resolution when analyzed with a single period window.

A rectangular window one fundamental period long has a bandwidth[3] in the frequency domain equal to twice the fundamental frequency, and when centered over a particular harmonic, its zero crossings occur exactly at the two adjacent harmonics, passing only the particular harmonic. Other extended FS windows have similar characteristics. Thus any broadening of the bandwidth of harmonics from the FS line spectrum assumed in Fig. 1(a), such as indicated by the FT in Fig. 1(b), causes leakage into the particular harmonic under analysis. The STFT and phase vocoder allow the use of longer windows, with narrower bandwidths to reduce leakage, but require time durations which may be long relative to audible temporal changes. The trade-off between window length and bandwidth is due to the minimum time-bandwidth product of the windows used in STFT's and extended FS methods, referred to as a version of the uncertainty principle[4] [44], [45].

STFT's and the related phase vocoder and extended FS methods are all represented in Cohen's class by their squared magnitudes. The kernel for computing these TF distributions using (2) is the WD $W_h(t,\omega)$ of the window function $h(t)$ [11] or

$$\varphi(t,\omega) = W_h(t,\omega). \tag{10}$$

Thus $h(t)$ is a rectangular window one period long for a pitch synchronous analysis, a Hanning window two periods long for Beauchamp's technique, and includes a much broader class of windows for the STFT or phase vocoder. In the case of the extended FS methods, the frequency values at which the magnitude is computed using (2) are restricted to multiples of the fundamental frequency $\omega = n\omega_0$. The WD of a signal satisfies its marginals, and thus when the kernel in (10) is used in (2), since the resulting linear transformation of the signal's WD spreads the signal energy over a larger region of the resulting TF distribution, reflecting the uncertainty principle requirements, the marginals are no longer satisfied. Therefore such time-window based distributions never satisfy both of their marginals.

Fig. 2(a) and (b) illustrate the application of an extended FS technique in two steps to our trumpet sample, showing the 16th–19th harmonics between the 80–400 ms points, spanning the duration of the attack. The magnitude squared of a STFT using a Hamming window two periods long (10.8 ms) of the average 185 Hz fundamental frequency is calculated first and shown in Fig. 2(a). The bandwidth of the extended FS window used in this case is 370 Hz, or two harmonic intervals, and because the components do not form a line spectrum, as shown in Fig. 1(b), the window creates leakage between them. The FS magnitude values versus time are obtained by sampling this distribution in frequency at multiples of 185 Hz, as shown in Fig. 2(b). We will see that these components actually change frequency

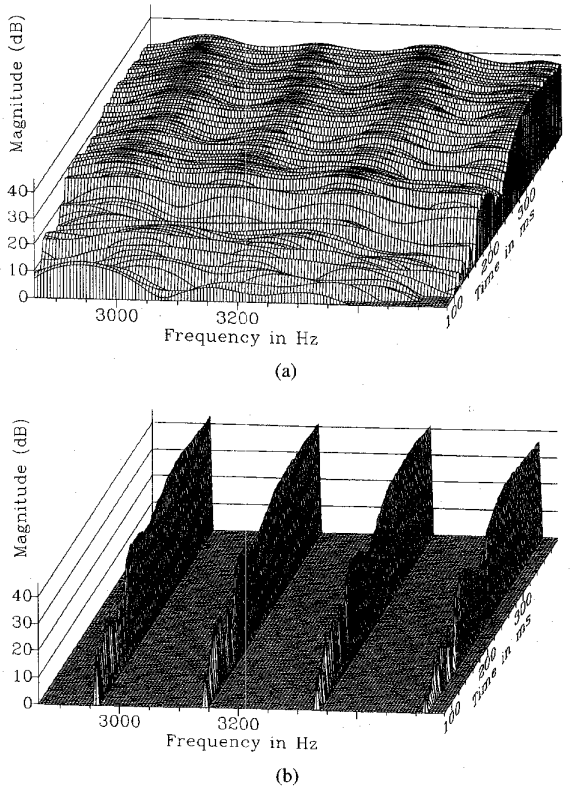


Fig. 2. Extended FS analysis applied to 16th–19th components of a trumpet attack (same signal as Fig. 1) over 300 ms time span, shown in two stages. (a) Application of the extended FS window showing frequency averaging due to wide bandwidth. (b) Harmonic interval sampling of (a) for final result, which does not track true component frequencies, and therefore distorts magnitudes.

over time, and thus this fixed sampling results in biased estimates of magnitude.

Fig. 3(a) illustrates the application of the more general STFT case to the same time and frequency range of our trumpet sample as Fig. 2(a) and (b). A longer, 188 ms Hamming window is utilized, whose narrower bandwidth avoids leakage between components. The STFT is also computed at a finer sampling of frequencies than the extended FS techniques, avoiding the bias of magnitude estimates in Fig. 2(b) which result from failure to follow the frequency shifts that are apparent in Fig. 3(a). However, the longer window averages the attack over a duration which is significant relative to the temporal resolution of human hearing and to the rate of change of this signal, which generates bias in estimating both magnitude and frequency.

A variation of the STFT which changes the window length as a function of frequency so that a constant number of periods are within the window at each frequency is the constant-Q distribution. Youngberg [46] explored the constant-Q transform for general acoustic signal analysis. Brown [47] applied the constant-Q transform to music, motivated by the fact that the geometrically distributed frequency resolution that it provides mirrors the structure of musical scales. The constant-Q transform is computed by contracting the STFT window in (9) with frequency to

[3]This is the bandwidth of the main lobe of the rectangular window's spectrum [14].

[4]The uncertainty principle has no probabilistic interpretation here, and simply describes how window bandwidth varies inversely with length, where a certain product of these factors has a fixed minimum.

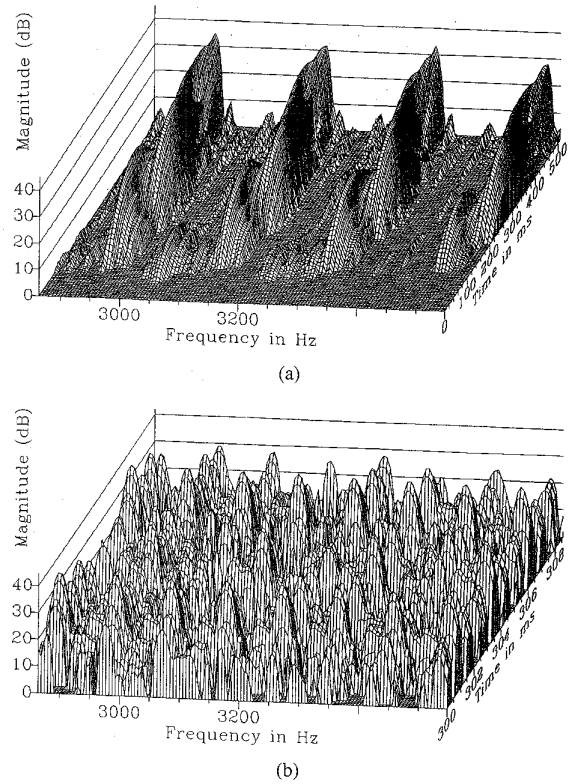


Fig. 3. Comparison of smoothing extremes of standard techniques applied to 16th–19th components of a trumpet attack (same signal as previous figures). (a) STFT of first 500 ms, showing smoothing and resulting bias required by the uncertainty principle. (b) PWD of 8 ms from the middle of the attack, with no smoothing in time, giving high resolution but allowing cross products to remain.

match the contraction of the complex exponential as given by

$$S^{CQ}(t,\omega) = \frac{1}{2\pi}\int_{-\infty}^{\infty} s(\xi)h\left((t-\xi)\frac{\omega}{\omega_0}\right)\exp(-j\omega\xi)\,d\xi \tag{11}$$

where ω_0 is a reference frequency at which the window is unaltered. The squared magnitude of the constant-Q transform is obtained in Cohen's class by using a frequency scaled window $h(t(\omega/\omega_0))$ to produce a kernel which is explicitly frequency dependent via (10), a variation on the STFT case. While this produces a constant-Q transform which has a local bandwidth proportional to frequency, providing a better fit for some applications by paralleling their structure, it remains bound by the uncertainty principle like any other time windowing method.

Equation (11) for the constant-Q distribution is also a type of wavelet transform. Daubechies' [48] research laid much of the theoretical foundation for wavelet transforms, which are obtained by decomposing signals into dilations and translations of a mother wavelet. If the window and complex exponential in (11) are combined for a particular frequency ω_m, a useful wavelet for musical signal analysis is obtained

$$w\left(\frac{t}{a}\right) = h\left(\frac{t}{a}\right)\exp\left(-j\frac{t}{a}\omega_m\right). \tag{12}$$

The scale variable a used to dilate the wavelet is inversely related to frequency ω ($a = 1$ generates the mother wavelet). More general types of wavelets are possible than that in (12). Various authors [49]–[51] have applied wavelets to the analysis, synthesis, and processing of musical sounds.

C. Generalized TF Distributions

The other fundamental line of research leading to TF distributions for music is that represented by Wigner *et al.* All of the TF distributions described above which adapt time-independent transforms to time-varying signals by using sliding windows assume that the signal is stationary over the window length. The transforms determine the spectral content within the window by decomposition into infinite duration, time-invariant sinusoids which do not reflect the true nature of a time-varying signal. Gabor [7] proposed a concept of instantaneous frequency, $\omega^i(t)$, and time varying amplitude or envelope $a(t)$ for a signal $s(t)$ where

$$s(t) = a(t)\cos(\phi(t)) \tag{13}$$

and

$$\omega^i(t) = \frac{d}{dt}\phi(t). \tag{14}$$

Equation (13) reduces to the constant parameter case $s(t) = a\cos(\omega t)$ if $a(t) = a$ and $\phi(t) = \omega t$, and thus is a generalization of the infinite duration sinusoidal concept of frequency and amplitude.

Gabor proposed using the Hilbert transform to resolve the ambiguity of how to decompose an arbitrary $s(t)$ into functions $a(t)$ and $\cos(\phi(t))$ to obtain instantaneous frequency and amplitude as given in (13) and (14). The Hilbert transform isolates the strictly positive frequencies in the signal, creating the complex analytic signal as given by

$$s^A(t) = a(t)\exp(\phi(t)). \tag{15}$$

Computing the instantaneous power $|s^A(t)|^2$ in this analytic signal gives $|a(t)|^2$, extracting the magnitude of the envelope from (15). The phase $\phi(t)$ is computed as $\arg(s^A(t))$, and $\omega_i(t)$ is computed from the phase using (14). The details of this are beyond the scope of this paper, but can be found in Boashash [52] and Picinbono [53].

The WD has several attractive properties related to Gabor's concepts of the analytic signal [17]. Because the WD satisfies its marginals, applying (6) to it expresses both the instantaneous power and the squared modulus of the complex envelope of the analytic signal. Furthermore, the average frequency of the WD is equal to the instantaneous frequency of the analytic signal [17]. These computations derive the instantaneous frequency and amplitude or envelope for the signal as a whole, but can be extended to

multiple individual components with reasonable restrictions on the components of a musical signal model [54], [55].

While the WD has many desirable properties as described above, it also presents difficulties to musical signal analysis. In Cohen's class the kernel that reproduces the WD from (2) is an impulse function, which imposes no averaging or smoothing at all, providing high resolution in time and frequency, but with a penalty. The bilinear character of the WD reflects the nonlinear nature of the computation of energy for all members of Cohen's class. These computations produce cross products between different signal components which are typically smoothed over by time-windowing kernels such as (10), but remain apparent in the WD, often introducing negative values. These cross products are illustrated with our trumpet sample in Fig. 3(b), where the pseudo-Wigner distribution (PWD) is shown, a variation of the WD used with finite data sets. This covers the same four harmonic frequency range as Figs. 1(b), 2(a), 2(b), and 3(a), but a time span of only 8 ms, because the WD's lack of smoothing requires a high display sampling rate to avoid aliasing.[5] Note the presence of cross products more widely distributed in the spectrum than the 185 Hz harmonic intervals seen in previous figures.

The flexibility of Cohen's class allows the design of a kernel for musical signals called the *modal* kernel, which can generate distributions between the STFT and the Wigner cases, reducing the smoothing bias of the STFT without incurring the cross products in the WD [54]–[57]. The design of the MD differs from that of many of the new generalized TF distributions in Cohen's class, such as reduced interference distributions (RID) [58] and cone-shaped kernel distributions [59]. The latter are optimized for properties which are desirable in the general theory of distributions, such as satisfaction of the marginals and positivity. The MD, in contrast, focuses on the application of estimating the instantaneous frequencies and amplitudes of a multicomponent musical signal model with minimum bias, where positivity is not important.

Key criteria for a musical signal distribution are wide dynamic range, frequency estimation accuracy on the order of cents,[6] averaging over time which is sufficiently minimized to retain features on the order of tens of milliseconds long, and low computational load. An important property in the estimation of the multicomponent signals like our trumpet example is limited superposition, which allows new components to be added without affecting the estimation of those already there if the new ones do not overlap the others. Riley [60] describes this concept, which prohibits cross terms between components, for speech processing with Cohen's class distributions.

The basis for the MD design is provided by the musical signal model

$$\overline{s}(t) = \sum_{l=1}^{M} \overline{B}_l(t) \exp(j\phi_l(t)) \tag{16}$$

where M positive frequency components are present, each having time-varying amplitudes and phases (allowing time-varying frequency). Each of the components must individually be an analytic signal, and components cannot overlap in the TF plane. This is a much less restricted model than extended FS methods allow. Considerable amplitude and frequency modulation is permitted, as are inharmonic components and multisource signals. The model in (16) is the basis for additive music synthesis [61], and resynthesis is directly facilitated by estimation of the model parameters.

The modal kernel is defined for Cohen's class by

$$\varphi_M(\tau, \xi) = h_{LP}(\tau) G_{LP}(\xi) \tag{17}$$

where $h_{LP}(\tau)$ and $G_{LP}(\xi)$ are low-pass filtering functions whose cutoff frequencies are directly related to signal properties such as the intercomponent frequency spacing. The averaging of the WD obtained when (17) in inserted in (2) is independently determined for time and frequency smoothing by the two factors in (17), which makes this a separable kernel. These independent factors allow smoothing with a time-bandwidth product less than those imposed by methods related to STFT's, including extended FS and constant-Q methods, while maintaining enough smoothing to suppress cross products, minimizing the overall bias in the distribution. The separable kernel in (17) also contributes to computational efficiency [55].

The modal kernel in (17) is related to the smoothed PWD [62]. The two were developed independently, and both are based on a simple time-smoothing of the PWD. The smoothed PWD was presented as a spectral estimator for a large class of random processes [63]. The MD was developed for estimating the parameters of the deterministic musical signal model in (16) [56]. To this end, favorable lowpass filter characteristics for $h_{LP}(\tau)$ and $G_{LP}(\xi)$ have been determined for the MD in that application. These are related to the signal-to-noise ratio (SNR) requirements for musical signal representation by the MD, and to the accuracy of new instantaneous frequency and power estimators developed specifically to operate on the discrete MD to estimate the continuous parameters in (16) [54], [55]. The lowpass filter cutoffs for $h_{LP}(\tau)$ and $G_{LP}(\xi)$ have been determined for a wide variety of signal types of specific interest in music. These cutoffs are based on the intercomponent frequency spacing and modulation of sinusoidally and exponentially amplitude modulated components, as well as sinusoidally and linearly frequency modulated components [55]. These cases cover the most common examples of musical signal component attack and decay characteristics, frequency and amplitude vibrato, and glissandi.

The MD is illustrated in Fig. 4(a) for our trumpet sample, which covers the same frequency range as Figs. 1(b), 2(a) and (b), and 3(a) and (b). The averaging over time is comparable to the favorable time resolution of the extended FS example in Fig. 2(a), but without the wide bandwidth which causes intercomponent contamination, or the problem of sampling at only harmonic frequencies of

[5] If a signal is sampled too sparsely relative to its rate of change, aliasing distorts the result [14].

[6] One cent is a frequency ratio of $2^{1/1200}$ = (one semitone)$^{1/100}$.

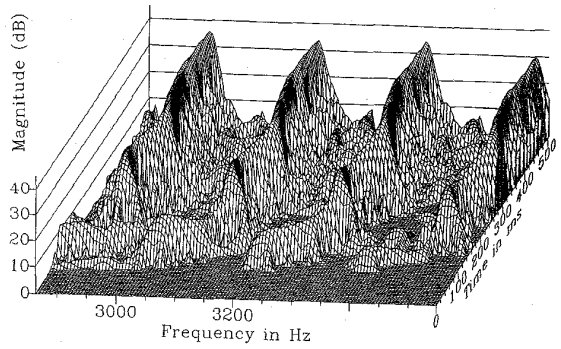


Fig. 4. MD's of trumpet attack (same signal as previous figures), combining the good time resolution of the extended FS analysis in Fig. 2 with very good and excellent frequency resolution. (a) 16th–19th components for comparison to previous figures, exhibiting half the time-bandwidth smoothing of extended FS or STFT methods. (b) 18th component, displayed with a time-bandwidth smoothing 23 times less than any extended FS or STFT method can achieve, but with none of the cross terms of the Wigner or PWD's.

Fig. 2(b). The averaging in frequency is comparable to the STFT in Fig. 3(a), where intercomponent contamination is eliminated by narrow bandwidth, without the additional time-averaging penalty in that figure. The reduced averaging is achieved without the cross term interference of the PWD in Fig. 3(b). Thus the MD captures the advantages of all these methods, without the disadvantages.

A number of interesting features of the trumpet sample are visible in Fig. 4(a). It is known [64] that the initial frequency for a trumpet note is set by the player with little help from the instrument, and that as the attack progresses, the player adjusts the lip frequency based on feedback from the instrument. These adjustments are multiplied with each harmonic in proportion to harmonic number. In the figure the 17th harmonic (second from the left) moves from an initial value of 2864 Hz to 3125 Hz in about 100 ms, which is a shift greater than a 185 Hz harmonic multiple. The rapid shifting in frequency as well as amplitude makes such an attack difficult to analyze by time windowing methods, which assume that each component remains relatively stationary. This component's representation in Fig. 4(a) as compared to Figs. 2(a) and 3(a) illustrates the distortions caused by window smoothing.

Fig. 4(a) also suggests that the amplitude shifts of the partials are highly coupled with the frequency shifts. In general, amplitude increments are associated with a (possibly short) period of frequency stability such that subsequent shifts in frequency may cause a drop in amplitude. These temporary peaks are the "blips" described earlier by various sources [15], [16], [35] and are clearly seen in Fig. 4(a) at the 19th (far right) component at about the 180 ms point. However, the extended FS methods used by other authors did not allow accurate tracking of amplitude versus frequency, because of the wide bandwidth associated with such windows. The extended FS plots in Figs. 2(a) and 2(b) show an amplitude peak around 180 ms, but obscures the associated frequency shifts.

While Fig. 4(a) represents a better than 2:1 reduction below the minimum time-bandwidth averaging of any window-based method, Fig. 4(b) shows the true power of the MD. Displayed is the 18th harmonic component of our trumpet sample [second from the right in Fig. 4(a)], with the time-bandwidth product of the kernel 23 times less than that achievable with any windowing method. Cross terms are still suppressed, showing that time-bandwidth product of the kernel can be reduced far below that required by window-based kernels without a cross term penalty.

IV. Extensions of Musical TF Analysis to Problems in Music Processing

Music processing shares many of the same applications as speech processing, e.g., analysis, synthesis, compression, and recognition. These are less developed, however, than other areas of music research, perhaps primarily because music processing has addressed historically the relatively focused needs of composers. Indeed, recently published edited volumes on music representation [65], [66] contain relatively little on what we have presented as the core of the signal representation problem, but investigate instead issues of representation at a "molar" level, such as one might see in a theoretical linguist's study of speech [67]. The following highlights several application areas and the impact that TF theory has had on them.

A. Analysis and Resynthesis

The composer and researcher J.-C. Risset performed an extensive analysis of trumpet tones [68]. He employed a succession of FFT's with a variable window length of 5–50 ms to capture the time-varying attributes of the timbre. Risset's research demonstrated that the spectrum of the trumpet is nearly harmonic, the higher harmonics become more prominent as the overall intensity increases, the amplitude envelope of the higher harmonics have a slower rise time and faster decay time than lower harmonics, the attack portion of the sound is characterized by small, quasi-random deviations in amplitude and frequency, and there is distinctive formant near 1500 Hz. Risset validated his analysis through software-based synthesis. The process of resynthesis not only confirmed the attributes discovered

in his research, but also confirmed the notion that certain characteristics of a time-varying spectrum play a key role in the perception of timbre.

Risset's work was extended by a series of historically important analyses of traditional acoustic instruments as codified in the Lexicon of Analyzed Tones [35]. Spectrographic plots of violin, clarinet, oboe, and trumpet tones depicted the changing amplitude of each partial over time. These analyses are an important avenue of study for sound synthesists interested in digitally recreating these particular instruments or altering characteristics of the plot to explore new timbres. Similarly, these analyses have been used to explore theoretical synthesis methodologies such as frequency modulation (FM) [69], waveshaping [70], [71], and more recent efforts in additive synthesis [72], [73], physical modeling [74], [75], and formant synthesis [76].

The advent of fast processors, and inexpensive storage and memory have made the power of software-based synthesis accessible to many musicians. For the first time in history, composers have an infinite palette of timbre available to them through software-based synthesis. The sheer magnitude of possible timbres and the myriad of parameters used to define those timbres beckons for intelligent methods to explore this expansive domain. Continued work in TF distributions of traditional acoustic instruments as well as synthetic sounds will assist composers in establishing an aesthetic framework for timbre.

These developments within music signal processing parallel those within speech processing where such tools as the spectrogram have played such an important role in understanding the features of the speech waveform. They differ, however, in the degree to which sophisticated TF techniques have been required, primarily because of the difference in focus on the signal. On the one hand, parametric modeling of the speech signal has sufficed in characterizing the gross features of the acoustic signal responsible for the basic phonetic component of speech communication. On the other hand, the high-fidelity requirements of the music signal have called for ever increasing refinements of the TF representation to reveal the signal dynamics that are responsible for our perception of the attack, decay, sustain, and release phases of something as simple as a musical note.

B. Transcription

Like its counterpart in speech processing, automatic music transcription presents a challenge that has required considerable research effort to obtain even the relatively few and meager successes of the past decade. Music transcription is complicated by the fact that, unlike the current demands of speech processing, the nuance of the performance (or utterance) is as much of interest as the musical score (or sequence of words) that is being performed. Complicating this fact is that most musical performances demand conditions that are known to compromise the best of speech recognition algorithms: the signal is made up of many voices or timbres, rather than just one, and recordings are often made under conditions of considerable multipath distortion (e.g., reverberation).

A base-level automated music transcription system takes a musical signal as input and from that signal, determines the fundamental pitch. The duration of the fundamental is placed within the context of time which renders this essential information as traditional music notation. Several approaches to this problem have led to successful single-timbre transcription systems including those by Piszczalski [77] and Moorer [33]. Within Moorer's approach, two voices of similar timbre and limited time variation have been transcribed under certain limiting conditions, but it appears that all approaches taken thus far are very sensitive to the presence of more than one particular timbre.

A variant on the transcription of pitch is the transcription of tempo or rhythm. Schloss' work [78] pertains to the transcription of percussive instruments where pitch is a more difficult attribute to define. This is related in scope to research on real-time automatic accompaniment systems [78] in which the computer accompanist must determine the tempo and microvariations of the soloist.

A more complete transcription system would not only determine the musical score performed by multitimbre sources, such as a symphonic orchestra, Gamelan ensemble, or jazz quartet, but would be able to transcribe the subtle (and not so subtle) expressiveness of the musicians in their performance. Such variations underly what is commonly assumed to distinguish a great from an average performance and would involve the transcription of microvariations in such gross features as tempo, intensity, and pitch, as well as the finer variations in timbre that reflect the different ways in which the performer generates the sound through their musical instrument. As Brown [47] has argued, these aspects of performance and its transcription cannot be easily documented with standard Fourier techniques and require a more powerful suite of TF tools.

C. Visualization

An emerging application which is relatively unique to the field of music processing is what we call visualization, although in a basic sense this problem is common to the discipline of signal representation, in general, and shares features with visual training devices in speech communications, in particular. By visualization we mean the mapping of an acoustic signal into visual form in order to better highlight desired features of the signal over others. Since antiquity, musicians master the subtle nuances of musical interpretation and expression through an auditory mentoring relationship with a teacher. A volley of teacher performs, student listens and subsequently performs, characterizes the apprenticeship. Extending feedback mechanisms to include visual as well as auditory perception may radically alter the pedagogy of musical performance, lead to better rates of learning, improved efficiency in practice, and greater accuracy in the control of music production process.

Visualization is the youngest and least developed of the applications areas discussed, but it is by far the most demanding of the technology. The tools of visualization must build from those TF distributions and algorithms that

are found to be most useful in analysis and transcription. To provide the flexibility in pedagogy, these algorithms must be implemented in real-time: in today's terms, our own internal fast-algorithm implementation of the modal kernel achieves a ratio of 100:1 in processing-to-real-time on a 90-MHz Pentium architecture for signals sampled at 44.1 kHz [80]. Finally, the exact form of the mapping must be considered carefully. What the eye sees in a visual representation of a TF surface is not always directly related to what the ear hears [81], [82]. Research on auditory front ends has much to offer designers of visual trainers for the recoding of the TF surface into more meaningful psychological units, such as pitch and timbre [83], [84].

V. Summary

We have traced the major time and frequency analysis methods that have been applied to music processing, and described application areas for these methods, with the contributions of key researchers highlighted in the review. Analyses spanning FS and FT's, several extended FS methods, variations on the STFT and constant-Q transform, the WD, and the MD were all covered. Techniques were compared by placing them in the context of Cohen's class of generalized TF distributions, which computes signal energy as a TF function. This facilitated comparison of existing methods, insight into the nature of the problem, and the design of new approaches. The limitations of windowing methods and their reliance upon steady-state assumptions and infinite duration sinusoids to define frequency and amplitude was demonstrated. The analytic signal as a basis for defining instantaneous frequency and power was shown to surpass these limitations, and led to the WD's and MD's, with the latter optimized for estimating these parameters in a musical signal model. The tenor of these techniques was further illuminated by their application to a trumpet sample chosen for its TF content. The strengths and limitations of each approach in revealing perceptual attributes of music, especially pitch and timbre, were described.

Music signal processing begins with a primitive signal representation, such as the TF representations which were the focus of much of this paper, and from these, develops more extensive representations. Applications to analysis and resynthesis, transcription, and visualization were described. Analysis and resynthesis were compared to the speech processing case, where the music signal's nature demands greater representation fidelity than speech often does. Music transcription was similarly compared to speech transcription, where the complications of multiple sources and reverberation are the norm, in contrast to speech where they can usually be avoided. Visualization was described as an emerging application, with the potential to radically alter the pedagogy of musical performance and practice. This application asks the most of all of TF methodology by combining the demands of real-time analysis and transcription with the need to portray the results in meaningful sensory displays.

References

[1] D. M. Green, *An Introduction to Hearing.* Hillsdale, NJ: L. Erlbaum, 1976.

[2] F. Opolko and J. Wapnick, McGill Univ. Master Samples compact discs, Music Dept., McGill Univ., Montreal, Quebec, Canada, 1987–1989.

[3] J. Jeong, "Time-frequency signal analysis and synthesis algorithms," Ph.D. dissertation, Dept. EECS, Univ. Michigan, 1990.

[4] O. Rioul and P. Flandrin, "Time-scale energy distributions: A general class extending wavelet transforms," *IEEE Trans. Acoust. Speech, Signal Process.*, vol. 40, pp. 1746–1757, July 1992.

[5] E. P. Wigner, "On the quantum correction for thermodynamic equilibrium," *Phys. Rev.*, vol. 40, pp. 749–759, 1932.

[6] J. Ville, "Theorie et applications de la notion de signal analytique," *Cables et Transmission*, vol. 2A, pp. 61–74, 1948.

[7] D. Gabor, "Theory of communication," *Proc. IEE*, vol. 93, no. 3, pp. 429–457, 1946.

[8] C. H. Page, "Instantaneous power spectra," *J. Appl. Phys.*, vol. 23, pp. 103–106, 1952.

[9] L. Cohen, "Time-frequency distributions—A review," *Proc. IEEE*, vol. 77, pp. 941–981, July 1989.

[10] ——, "Generalized phase-space distribution functions," *J. Math. Phys.*, vol. 7, pp. 781–786, 1966.

[11] T. A. C. M. Claasen and W. F. G. Mecklenbrauker, "The Wigner distribution–A tool for time-frequency analysis, Part III: relations with other time-frequency signal transformations," *Phillips J. Res.*, vol. 35, no. 6, pp. 373–389, 1980.

[12] H. Helmholtz, *On the Sensations of Tone as a Physiological Basis for the Theory of Music,* A. J. Ellis, transl. New York: Dover, 1954.

[13] I. Gratton–Guiness, *Joseph Fourier 1768-1830.* Cambridge, MA: MIT Press, 1972.

[14] R. A. Roberts and C. T. Mullis, *Digital Signal Processing.* Reading, MA: Addison-Wesley, 1987.

[15] D. Luce, "Physical correlates of nonpercussive musical instruments," Ph.D. dissertation, EECS Dept., MIT, Cambridge, MA, 1963.

[16] J. C. Risset and M. Mathews "Analysis of musical-instrument tones," *Phys. Today*, vol. 22, no. 2, pp. 23–30, 1969.

[17] T. A. C. M. Claasen and W. F. G. Mecklenbrauker, "The Wigner distribution—A tool for time-frequency analysis, Part I: Continuous-time signals," *Phillips J. Res.*, vol. 35, no. 3, 217–250, 1980.

[18] J. M. Grey, "An exploration of musical timbre using computer-based techniques for analysis, synthesis, and perceptual scaling," Ph.D. dissertation, Dept. Psychol., Stanford Univ., 1975.

[19] E. Terhardt, "Frequency analysis and periodicity detection in the sensations of roughness and periodicity pitch," in *Frequency Analysis and Periodicity Detection in Hearing,* R. Plomp and G. F. Smoorenburg, Eds. Leiden, The Netherlands: Sijthoff, 1970, pp. 278–287.

[20] J. W. Beauchamp, "Electronic instrumentation for the synthesis, control, and analysis of harmonic musical tones," Ph.D. dissertation, Dept. Electrical Eng., Univ. Illinois, 1965.

[21] E. Terhardt, G. Stoll, and M. Seewann "Algorithm for extraction of pitch and pitch salience from complex tonal signals," *J. Acoust. Soc. Amer.*, vol. 71, no. 3, pp. 679–688, 1982.

[22] C. Seeger, "An instantaneous music notator," *J. Int. Folk Music Council*, vol. 3, pp. 103–106, 1951.

[23] M. Hood, *The Ethnomusicologist.* New York: McGraw-Hill, 1971.

[24] H. Fletcher, E. D. Blackham, and R. Stratton, "Quality of piano tones," *J. Acoust. Soc. Amer.*, vol. 34, no. 6, pp. 749–761, 1962.

[25] H. Fletcher, E. D. Blackham, and D. A. Christensen, "Quality of organ tones," *J. Acoust. Soc. Amer.*, vol. 35, no. 3, pp. 314–325, 1963.

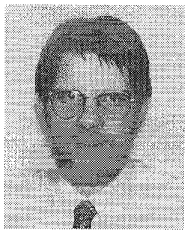
[26] H. Fletcher, E. D. Blackham, and O. N. Geertsen, "Quality of violin, viola, 'cello, and bass-viol tones. I," *J. Acoust. Soc. Amer.*, vol. 37, no. 5, pp. 851–863, 1965.

[27] M. V. Mathews, J. E. Miller, and E. E. David Jr., "Pitch synchronous analysis of voiced sounds," *J. Acoust. Soc. Amer.*, vol. 33, no. 2, pp. 179–186, 1961.

[28] D. Luce and M. Clark, Jr., "Physical correlates of brass-instrument tones," *J. Acoust. Soc. Amer.*, vol. 42, no. 6, pp. 1232–1243, 1967.

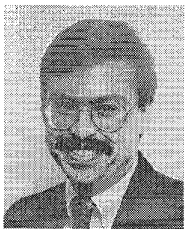
[29] W. Strong and M. Clark, "Synthesis of wind instrument tones," *J. Acoust. Soc. Amer.*, vol. 41, no. 39, 1967.
[30] J. C. Risset and M. Mathews, "Analysis of musical-instrument tones," *Phys. Today*, vol. 22, no. 2, pp. 23–30, 1969.
[31] M. V. Mathews and J. Kohut, "Electronic simulation of violin resonances," *J. Acoust. Soc. Amer.*, vol. 53, no. 6, pp. 1620–1626, 1973.
[32] J. W. Beauchamp, "Time-variant spectra of violin tones," *J. Acoust. Soc. Amer.*, vol. 56, no. 2, pp. 995–1004, 1974.
[33] J. A. Moorer, "On the segmentation and analysis of continuous musical sound by digital computer, Ph.D. dissertation, Dept. Computer Sci., Stanford Univ., Stanford, CA, 1975.
[34] J. M. Grey and J. A. Moorer, "Perceptual evaluations of synthesized musical instrument tones," *J. Acoust. Soc. Amer.*, vol. 62, no. 3, pp. 454–462, 1977.
[35] J. A. Moorer, J. M. Grey, and J. Strawn, "Lexicon of analyzed tones, Part 3: The trumpet," *Computer Music J.*, vol. 2, no. 2, pp. 23–31, 1977.
[36] J. L. Flanagan and R. M. Golden, "Phase vocoder," *Bell Syst. Tech. J.*, vol. 45, pp. 1493–1509, 1966.
[37] R. W. Shafer and L. R. Rabiner, "Design and simulation of a speech analysis-synthesis system based on short-time Fourier analysis," *IEEE Trans. Aud. and Electroacoust.*, vol. AU-21, no. 3, pp. 165–174, 1973.
[38] M. Portnoff, "Implementation of the digital phase vocoder using the fast Fourier transform," *IEEE Acoust., Speech, Signal Process.*, vol. ASSP-24, no. 3, pp. 243–248, 1976.
[39] ——, "Time-scale modification of speech based on short time Fourier analysis," Ph.D. dissertation, MIT, Cambridge, MA, 1978.
[40] ——, "Time-frequency representation of digital signals and systems based on short-time Fourier analysis," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-28, pp. 55–69, Jan. 1980.
[41] M. B. Dolson, "A tracking phase vocoder and its use in the analysis of ensemble sounds," Ph.D. dissertation, Calif. Inst. Technol., 1982.
[42] J. Strawn, "Modeling musical transitions," Ph.D. dissertation, Dept. Music, Stanford Univ., 1985.
[43] N. F. Viemeister and C. J. Plack, "Time analysis," in *Human Psychophysics,* W. Yost, A. Popper, and R. Fay, Eds. New York: Springer-Verlag, 1993
[44] R. L. Bracewell, *The Fourier Transform and Its Applications.* New York: McGraw-Hill, 1986.
[45] N. G. de Bruijn, "Uncertainty principles in Fourier analysis," in *Inequalities II,* O. Shidha, Ed. New York: Academic, 1967, pp. 57–71.
[46] J. E. Youngberg, *A Constant Percentage Bandwith Transform for Acoustic Signal Processing.* Salt Lake City: Univ. Utah, 1980.
[47] J. C. Brown, "Calculation of a constant-Q spectral transform," *J. Acoust. Soc. Amer.,* vol. 89, no. 1, pp. 425–434, 1991.
[48] I. Daubechies, "The wavelet transform, time-frequency localization, and signal analysis," *IEEE Trans. Inform. Theory,* vol. 36, no. 5, pp. 961–1005, 1990.
[49] G. Evangelista, "Wavelet transforms that we can play," in *Representations of Musical Sounds,* G. De Poli, A. Piccialli, and C. Roads, Eds. Cambridge, MA: MIT Press, 1991.
[50] R. Kronland–Martinet, "The wavelet transform for analysis, synthesis, and processing of speech and music sounds," *Computer Music J.*, vol. 12, no. 4, pp. 11–20, 1988.
[51] R. Kronland–Martinet and A. Grossmann, "Application of time-frequency and time-scale methods to the analysis, synthesis, and transformation of natural sounds," in *Representations of Musical Sounds*, G. De Poli, A. Piccialli, and C. Roads, Eds. Cambridge, MA: MIT Press, 1991.
[52] B. Boashash and G. Jones, "Instantaneous frequency and time-frequency distributions," in *Time-Frequency Signal Analysis,* B. Boashash, Ed. New York: Wiley, 1992.
[53] B. Picinbono, "The analytical signal and related problems," in *Time and Frequency Representations of Signals and Systems,* G. Longo and B. Picinbono, Eds. New York: Springer-Verlag, 1989.
[54] W. J. Pielemeier and G. H. Wakefield, "Multi-component power and frequency estimation for a discrete TFD," in *Proc. IEEE-SP Int. Symp. on Time-Frequency and Time-Scale Anal.*, Philadelphia, PA, 1994.
[55] W. J. Pielemeier, "A time-frequency distribution with parameter estimators for a musical signal model," Ph.D. dissertation, Elect. Engin. and Comp. Sci. Dept., Univ. Mich., 1994.
[56] W. J. Pielemeier and G. H. Wakefield, "Time-frequency and time-scale analysis for musical transcription," *Proc. IEEE-SP Int. Symp. on Time-Frequency and Time-Scale Anal.*, Victoria, BC, Canada, 1992.
[57] ——, "A high-resolution time-frequency representation for musical instrument signals," *J. Acoust. Soc. Amer.*, vol. 99, no. 4, 1996.
[58] J. Jeong and W. J. Williams, "Kernel design for reduced interference distributions," *IEEE Trans. Signal Process.*, vol. 40, no. 2, pp. 402–412, 1992.
[59] Y. Zhao, L. E. Atlas, and R. J. Marks II, "The use of cone shaped kernels for generalized time-frequency representations of nonstationary signals," *IEEE Acoust. Speech, Signal Process.*, vol. 38, 1084–1091, July 1990.
[60] M. D. Riley, *Speech Time-Frequency Representations.* Boston: Kluwer, 1989.
[61] C. Dodge and T. A. Jerse, *Computer Music: Synthesis, Composition and Performance.* New York: Schirmer, 1985.
[62] W. Martin and P. Flandrin, "Wigner–Ville spectral analysis of nonstationary processes," *IEEE Acoust. Speech, Signal Process.*, vol. ASSP-33, pp. 1461–1470, June 1985.
[63] ——, "Analysis of nonstationary processes: short time periodigrams vs. a pseudo Wigner estimator," in *Signal Processing II, Proc. EUSIPCO-83*, H. W. Schussler, Ed., Erlagen, Germany, Sept. 1983, pp. 455–458.
[64] N. H. Fletcher and T. D. Rossing, *The Physics of Musical Instruments.* New York: Springer-Verlag, 1991, pp. 384–392.
[65] P. Howell, R. West, and I. Cross, Eds., *Representing Musical Structure.* New York: Academic, 1991.
[66] G. De Poli, A. Piccialli, and C. Roads, Eds., *Representations of Musical Sounds.* Cambridge, MA: MIT Press, 1991.
[67] F. Lerdahl and R. Jackendoff, *A Generative Theory of Tonal Music.* Cambridge, MA: MIT Press, 1983.
[68] J. C. Risset, "Computer study of trumpet tones (with sound examples)," in *Bell Laboratories Report.* Murray Hill, NJ: Bell Labs, 1966.
[69] J. Chowning, "The synthesis of complex audio spectra by means of frequency modulation," *J. Audio Engin. Soc.*, vol. 21, no. 7, pp. 526–534, 1973.
[70] D. Arfib, "Digital synthesis of complex spectra by means of multiplication of nonlinear distorted sine waves," *J. Audio Engin. Soc.*, vol. 27, no. 10, pp. 757–768, 1979.
[71] M. LeBrun, "Digital waveshaping synthesis," *J. Audio Engin. Soc.*, vol. 27, no. 4, pp. 250–266, 1979.
[72] J. W. Beauchamp, "Synthesis by spectral amplitude and 'brightness' matching of analyzed musical instrument tones," *J. Audio Engin. Soc.*, vol. 30, no. 6, 1982.
[73] X. Serra, "A system for sound analysis/transformation/synthesis based on a deterministic plus stochastic decomposition," Ph.D. dissertation, Dept. Music, Stanford Univ., 1989.
[74] M. E. Mcintyre, R. T. Schumacher, and J. Woodhouse, "On the oscillations of musical instruments," *J. Acoust. Soc. Amer.*, vol. 74, no. 5, pp. 1325–1345, 1983.
[75] J. O. Smith, "Musical applications of digital waveguides," Rep. No. STAN-M-39, Dept. Music, Stanford Univ., 1987.
[76] X. Rodet, "Time-domain formant-wave function synthesis," *Comp. Music J.*, vol. 8, no. 3, pp. 9–14, 1985.
[77] M. Piszczalski, "A computational model of music transcription," Ph.D. dissertation, Dept. Comp. Sci. and Engin., Univ. Mich., 1986.
[78] A. W. Schloss, "On the automatic transcription of percussive music—From acoustic signal to high-level analysis," Ph.D. dissertation, Dept. Music, Stanford Univ., 1985.
[79] B. Vercoe, "The synthetic performer in the context of live performance," *Proc. 1984 Int. Comp. Music Conf.*, pp. 199–200, 1984.
[80] W. J. Pielemeier, R. Guevara, and G. H. Wakefield, "How to attack a piano attack—An application of time-frequency estimation," *J. Acoust. Soc. Amer.* vol. 95, no. 5, pt. 2, 1994.
[81] S. Scruggs and G. H. Wakefield, "Time-frequency representation of auditory signatures," in *1992 IEEE Digital Signal Process. Workshop*, 1992.
[82] S. Scruggs, "A measurement-based approach to signal theory," Ph.D. dissertation, Univ. Mich., 1993.
[83] R. Patterson *et al.*, "Complex sounds and auditory images," in *Auditory Physiology and Perception,* Y. Cazals, L. Demany, and L. Horner, Eds. Oxford, U.K.: Pergamon, 1992, pp. 429–443.

[84] M. Slaney and R. Lyon, "Apple hearing demo reel," Apple Tech. Rep. No. 25, Apple Computer Corp., Cupertino, CA, 1991.



William J. Pielemeier (Member, IEEE) received the B.F.A. degree in mathematics from Kalamazoo College, Kalamazoo, MI, in 1978, and the B.S., M.S., and Ph.D. degrees in electrical engineering from the University of Michigan, Ann Arbor, in 1984, 1989, and 1993, respectively.

From 1976 to 1979, he was with the Special Projects Section of Whirlpool Corporation, Benton Harbor, MI, where he worked on manufacturing techniques for thick film hybrid circuitry. From 1981 to 1989, he worked for Tech-S, Inc., as a digital design engineer. In 1989 he became a Research Assistant in the graduate program at the University of Michigan, where he designed high-speed trinary logic serial interfaces for micromachined neural microprobes, and conducted research on musical signal analysis and time-frequency distributions. In 1993 he joined Ford Research Laboratories, Dearborn, MI, where he is currently applying signal processing and psychophysical testing to the study of human vibration perception. He continues to work independently on time-frequency estimation for musical signals and transcription.



Gregory H. Wakefield (Member, IEEE) received the B.A. degree (summa cum laude) in mathematics and psychology, the M.S. and Ph.D. degrees in electrical engineering, and the Ph.D. degree in psychology from the University of Minnesota, Minneapolis, in 1978, 1982, 1985, and 1988, respectively.

In 1986 he joined the faculty of the University of Michigan's Electrical Engineering and Computer Science Department, where he is currently an Associate Professor. He also serves as the Graduate Chair of the EECS–Systems Science and Engineering Division. His research interests are drawn from statistical signal processing, music signal processing, auditory systems modeling, psychoacoustics, sensory prosthetics, and sound quality engineering. He serves as a consultant to the Scientific Research Laboratories of Ford Motor Company and has consulted for 3M and Honeywell.

Dr. Wakefield received the NSF Presidential Young Investigator Award in 1987. He is a member of Phi Beta Kappa and Sigma Xi.

Mary H. Simoni received the B.A., M.A., and Ph.D. degrees in theory and composition, instrumental school music, and music theory from Michigan State University, East Lansing. She also studied at Stanford University's Center for Computer Research in Music and Acoustics and at the Center for Computer Music, City University of New York.

She is presently Assistant Professor of Music and Computer Technology at the University of Michigan, Ann Arbor. She is also Director of the university's Center for Performing Arts Technology and Chair of the Department of Media and Music Technology. She has taught at Berklee College of Music, Boston, MA, Lansing Community College, and Michigan State University. She also served as Supervisor of public computing operations at the University of Michigan. From 1990 to 1992 she was one of 12 scholars selected nationally to explore the use of multimedia to teach concepts in the liberal arts.