A SPECTROGRAM FOR THE TWENTY-FIRST CENTURY

Sean A. Fulop
Department of Linguistics, California State University, Fresno
Fresno, California 93740

Kelly Fitz
School of Electrical Engineering and Computer Science, Washington State University
Pullman, Washington 99164

Just as World War II was breaking out in Europe in 1939, a prototype of a remarkable electrical device was being completed at Bell Telephone Laboratories, under the direction of Ralph Potter. This device was able to provide, on a strip of paper, a continuous running document of the Fourier spectrum of a sound signal as it changed through time. Because of the war it was kept under wraps, but its detailed construction and numerous applications were revealed to the scientific community in a series of papers published in the *Journal of the Acoustical Society of America* (JASA) in 1946,[1,2] wherein it was called the Sound Spectrograph. The running spectral analysis that it output was termed a *spectrogram*.

The spectrograph has been recorded in history as one of the most useful and influential instruments for acoustic signal processing. In particular, the fields of phonetics and speech communication, which motivated the development of the machine, have been completely transformed by its widespread adoption. Over the decades, the cumbersome and delicate analog spectrograph hardware was transformed into more robust digital hardware at first, and then as computers became generally more powerful, into the digital software incarnations most of us use today. The underlying principle of the spectrogram has never changed; most applied acousticians who do time-frequency analysis are content to use software that in essence simulates the output that appeared 60 years ago in *JASA* (Fig. 1). Of what else in acoustics can the same be said? Do we use 60-year old microphones? Tape recorders? Loudspeakers?

Well, in truth, some of us have not been so content, but a more useful analytical process has never been generally recognized. The rich area of signal pro-
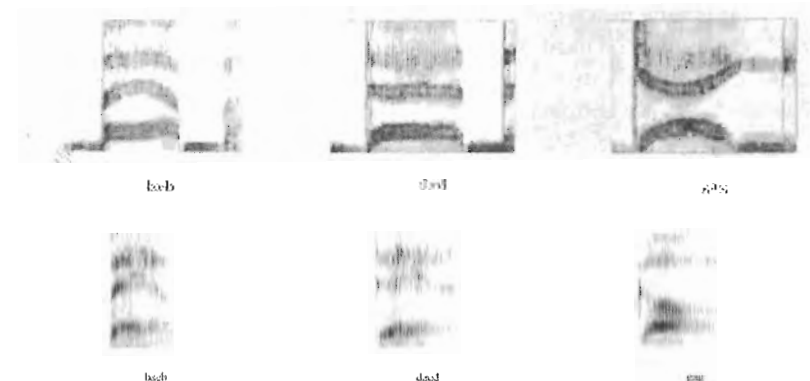

Fig. 1. Compare spectrograms of utterances [baeb], [daed], [gag] from a 1946 JASA paper[2] (upper panel) to ones made in 2006 from the first author's similar utterances, using the popular Praat sound analysis software.

cessing research known as "time-frequency analysis" grew up during the past 60 years. But the numerous variations on the spectrogram that have been touted (one can think of the Wigner-Ville transform, or wavelet analysis) have never made much impact in many applied circles because of physical interpretation problems, readability problems, or simply because they were not that much better. Time-frequency representations other than the spectrogram, while they may be more precise for certain test signals like a pure chirp (frequency-modulated sinusoid), usually provide unwanted "cross-terms" that do not correspond to physically or auditorily interpretable sound (signal) components. Ones that do not have this problem do not look much better than a spectrogram in any case. Wavelet processing is better thought of as "time-scale" rather than "time-frequency," and yields representations which are hard to read and interpret in spectrographic terms.

In fact, one especially useful improvement on the spectrogram (for many applied purposes, at least) was devised and implemented as software thirty years ago, but it was somehow drowned in the signal processing hub-

bub. Too many hopeful but ultimately useless efforts at time-frequency analysis left everybody jaded and defeatist, and it has taken these thirty years for the interest, invention, and mathematical analysis of a number of researchers to bring us to a point where we can well and truly say that there is something out there which is better than a spectrogram. The *reassigned spectrogram* is ready for its close-up.

Fourier's timeless problem

The time-frequency analysis of a signal refers generally to a three-dimensional representation showing the passage of time on one axis, the range of frequencies on a second axis, and the amplitude found in each time-frequency intersection (or *cell* in the digital domain) on a third axis. The amplitude axis is traditionally shown by linking the values to a grayscale colormap over a two-dimensional *time-frequency matrix*. This kind of time-frequency representation attempts to show the distribution of signal energy over the time-frequency plane. The archetypal time-frequency representation is the spectrogram, which was originally developed using analog electrical filters, but which was eventually

represented mathematically as the magnitude of the short-time Fourier transform,[3] a two-dimensional time-varying generalization of the Fourier transform.

Briefly put, for each time step of a signal, a spectrogram (Fig. 2) shows the decomposition of an analysis window into its Fourier spectral components as defined using the conventional Fourier transform.

This analysis produces a problem resulting from the inappropriateness of the Fourier transform's formalization of the colloquial notion of "frequency" when a short analysis span is used in the spectrogram. A mathematical frequency defined using the Fourier transform does not correspond to our intuitive understanding of "frequency" unless the analysis span is infinite in time, and this negates the ability of a time-varying generalization of the scheme to tell us what we want to know about each short frame. For example, the Fourier spectrum of frequencies in a sine wave of infinite extent is indeed just the frequency of the sine wave which we would intuitively want (meaning, it is just exactly the back-and-forth rate of the oscillation), but when the sine wave is not infinite-time, the Fourier spectrum instead yields a band
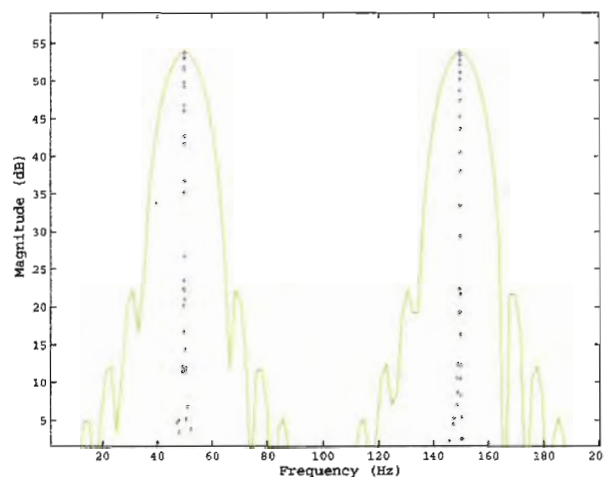

Fig. 3. *Fourier spectrum (green line) versus reassigned instantaneous frequency spectrum (blue points) of a two-component signal, both computed from a single Fourier transform over a 125 ms frame. The precise location of the 50 and 150 Hz sinusoidal components is shown by computing their instantaneous frequencies and reassigning points, but not by means of the Fourier frequency definition.*

Fig. 2. *Conventional (upper) and reassigned spectrograms of a pure double chirp signal, both computed using 15.6 ms frames, 156 μsec frame overlap. The reassigned spectrogram displays amplitudes from red (loud) to blue (quiet) in spectral order. The increased precision of frequency tracking, is notable, in spite of visible computational artifacts and interference lines.*

of frequencies surrounding and obscuring the intuitive frequency of the sine wave (see Fig. 3).

This so-called "smearing" in frequency affects the conventional spectrogram of Fig. 2, which is comprised of a sequence of Fourier spectra of successive short analysis windows of the signal. A mathematical duality within the transform induces a corresponding smearing in time, which may serve to obscure the true times of excitation of the various frequencies.

It is important to recognize that the frequency smearing that is so egregious in a short frame spectrogram is *not* a result of the uncertainty principle (shown by Denis Gabor[4] to be analogous to the Heisenberg principle) which governs the duality between time and frequency, as this affects the *resolving power* of the transform in the time and frequency dimensions. The smearing is a *precision* problem rather than a resolution problem, and this is clear from the fact that even one purely sinusoidal signal component will be smeared in a conventional wideband spectrogram, whether or not we attempt to resolve it from anything else.

In speech analysis and many other applications, the investigator is frequently not interested in the time-frequency energy distribution that the spectrogram provides, but is rather more interested in the *instantaneous frequencies* of the various amplitude-modulated (AM) or frequency-modulated (FM) sinusoidal components (often called line components) of a multicomponent signal. The instantaneous frequency[5,6] is a suitable generalization of mathematical frequency that may change over time. Specifically, it is defined as the derivative of the frequency modulation function of a single line component—this degenerates formally to the intuitive frequency of an unmodulated sine wave, no matter how long or short that sine wave is.

The reason to switch our mathematical model of intuitive "frequency" from Fourier's definition to that of instantaneous frequency is chiefly this: It seems increasingly likely that the human auditory system somehow pays attention to instantaneous frequency rather than the classical Fourier frequency. First, no one has reported that the auditory per-
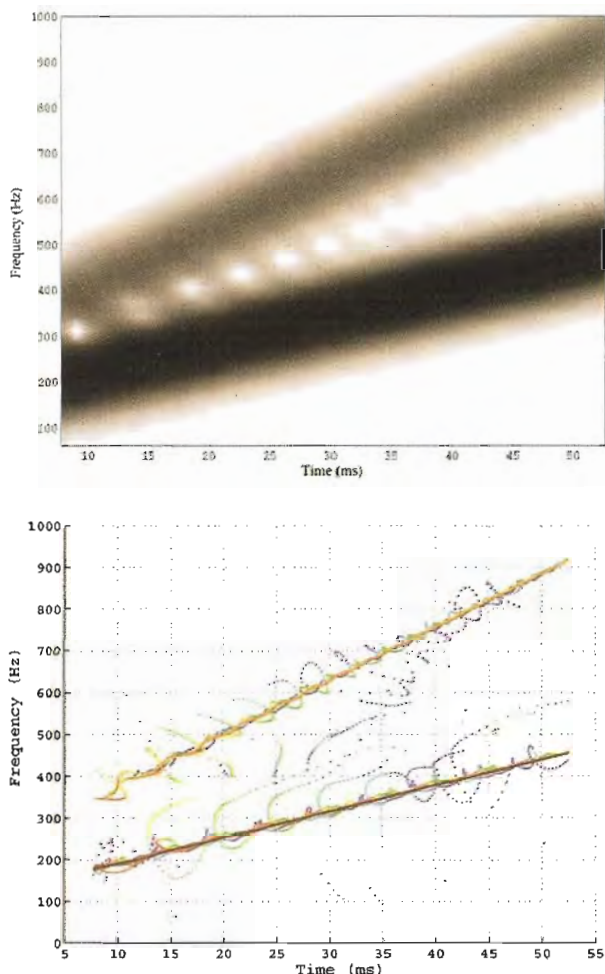
ception of the spectral timbre of a single unmodulated component (sine wave) is different depending on how long the signal was played, so long as the length of time exceeds a minimum auditory threshold of timbre perception. Additionally, the human auditory system has the ability to track the changing pitch of a frequency modulated sinusoid over time, while retaining the perception of sine wave timbre just as if there were no frequency modulation. Neither of these attributes is true of the classical Fourier spectrum of such sounds—a short sine wave supposedly has a larger and louder band of extra frequency components than a longer one, and a chirp has a similar band of such "components" surrounding the single modulated one we know is there. So, the timbre analysis that is performed by the ear and brain is probably not like a Fourier analysis, since we do not auditorily experience the spectral smearing of frequencies when they are modulated.

Components, by any other frame, would sound complete

The conventional spectrogram, as with all so-called *time-frequency representations*, provides us with a picture of how the signal's energy is distributed in time and frequency—Fourier's frequency, that is. This is the root of its problems. Rather than trying to improve this representation, which decades of signal processing research have shown to be impossible to any significant extent, a few independent thinkers worked to develop a new kind of image that would show the time course of the instantaneous frequencies of the components in a multicomponent signal. The seminal papers which put forth the original technique—called the *modified moving window method*—were written by Kunihiko Kodera and his colleagues C. de Villedary and R. Gendrin.[7,8] Their jumping off point was a paper published by Rihaczek[9] which demonstrated the connection between instantaneous frequency and the argument (phase) of a *complex analytic* signal—basically a complex version of a real signal, where the nonphysical imaginary part is related by the Hilbert transform to the physical real part. The main insight added by Kodera *et al.* was the recognition that the short-time Fourier transform—a 2D function having complex values whose magnitude yields the spectrogram—can be regarded as a "channelizing" of the signal into a bunch of complex analytic signals, one for each Fourier spectral frequency, from which the instantaneous frequencies of line components in the signal can be computed.

Unfortunately their method was almost perfectly ignored by the signal processing and acoustics communities at that time and for many years afterward, an all-too-common attribute of truly original work. More recently, however, the idea has been revived through alternative methods,[10,11] and also followed up with theoretical improvements,[12] and these developments have led to the adoption of the reassigned (also more wordily called the *time-corrected*

"One especially useful improvement on the spectrogram was devised and implemented as s oftware thirty years ago, but it was somehow drowned in the signal processing hubbub."

instantaneous frequency) spectrogram by a small number of applied researchers.

To understand the new approach, one must consider the signal as a superposition, not of pure sine waves as Fourier taught us, but rather of the generalized line components already mentioned, which may have amplitude or frequency modulation. The objective now is to compute the instantaneous frequencies of these line components as the signal progresses through time. Because of a duality within the procedure, it is also possible to compute a sort of "instantaneous time point" for the excitation of each line component along the way, which is technically the group delay associated with each digital time index.

Note that since **there is no mathematically unique way to decompose a signal into line components**, the decomposition always depends upon the analysis frame length, with longer frames able to capture lower-frequency components and also to resolve multiple components which are close in frequency. Yet, each different decomposition of a signal is an equally valid representation of its physical nature—it is up to the analyst to decide whether a long or short frame analysis is appropriate to highlight the physical aspects that are of interest in a particular case.

As Kodera *et al.* demonstrated, the partial derivative in time of the complex short-time Fourier transform (STFT) argument defines the *channelized instantaneous frequencies*, so that in the digital domain each quantized frequency bin is assumed to contain, at most, one line component, and the instantaneous frequency of that component is thus computed as the time derivative of the complex angle in that frequency bin.[10] Dually, the partial derivative in frequency of the STFT phase defines the *local group delay*, which provides a time correction that pinpoints the precise occurrence time of the excitation of each of the line components.[10] In a reassigned spectrogram (e.g., Figs. 2 and 4) the computed line components are plotted on the time-frequency axes, with the magnitude of the STFT providing the third dimension just as in the conventional spectrogram. One literally *reassigns* the time-frequency location of each point in the spectrogram to a new location given by the channelized instantaneous frequency and local group delay, whereas in the conventional spectrogram the points are plotted on a simple grid, at the locations of the Fourier frequencies and the time indices. A by product of this is that the reassigned spectrogram is no longer a time-frequency representation, nor even the graph of a function; the images presented here are 3D scatterplots with their z-axis values shown by a colormap. This article will not delve into the particular algorithms that may be used to compute a reassigned spectrogram—the problem there is in essence how one computes the time and frequency derivatives of the complex STFT argument. An historical and technical review of the subject, complete with a variety of algorithms for computation, is provided elsewhere by the authors.[13]

Applications of the reassigned spectrogram

The above explanations show that a goodly amount of reconceptualizing and unfamiliar new mathematical values are involved in the reassigned spectrogram, and applied researchers are going to want a big payoff in order to find all this worth while. Are we seriously going to toss aside the principles of Fourier analysis in favor of something new? The proof of the pudding is in the eating, so the chief justification for doing so is that we will then be better able to see and measure quantities of interest.

Speech sound analysis

One of the most important application areas of this technology is the imaging and measurement of speech sound. It has been found to be particularly useful for analyzing the fine scale time-frequency features of individual pulsations of the vocal cords during phonation. The phonation process involves the repetitive acoustic excitation of the vocal tract air chambers by the periodic release of air puffs by the vocal cords, which in an idealized model provide spectrally tilted impulses. As Fig. 4 shows, the reassigned spectrogram provides an impressive pic-

ture of the resonance frequencies (known as *formants*) and their amplitudes following excitation by each such impulse.

The speech signal in Fig. 4 was produced using *creaky* phonation, which has a very low frequency and airflow, and is thus quite pulsatile. This is to eliminate aeroacoustic effects and render the process as purely acoustic as possible for illustration.

Under more natural phonatory conditions, the process is significantly aeroacoustic, meaning that the higher airflow of ordinary speech cannot be neglected and has clearly observable effects on the excitation of resonances. The complexity of real phonation often militates against the measurement of formants, but we hope that the reassigned spectrogram can make such measurement possible, even easy.

Fishy signals

Gymnotiform fish of the tropical Americas and Africa produce quasi-sinusoidal electrical discharges of low amplitude by means of a neurogenic organ.[14] While not acoustic in nature, the signals are now believed to have a communicative function, and so the effective analysis of such signals poses a problem that is very much in the vein of bioacoustic signal processing.
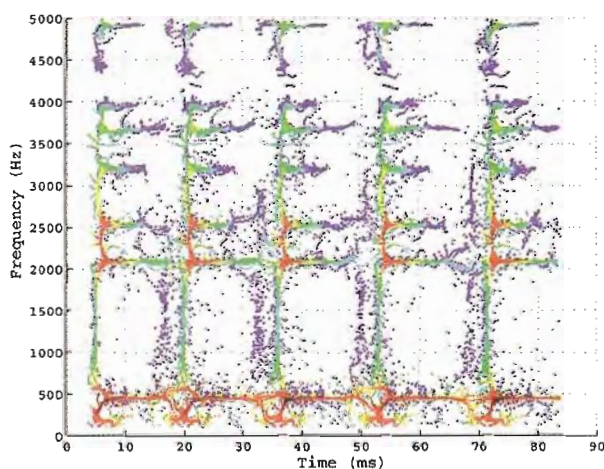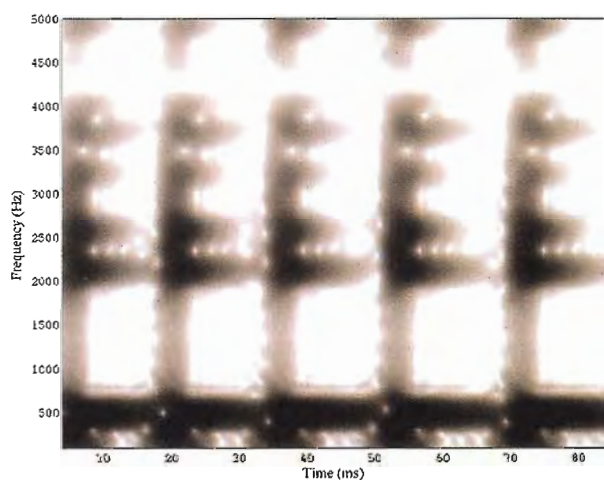

Fig. 4. Conventional and reassigned spectrograms of several vocal cord pulsations during the vowel [e] 'hay' pronounced with creaky phonation; computed using 7.8 ms analysis frames. Vocal tract resonances appear as red or yellow excitations at each impulse, and they decay for some time before the next impulse.
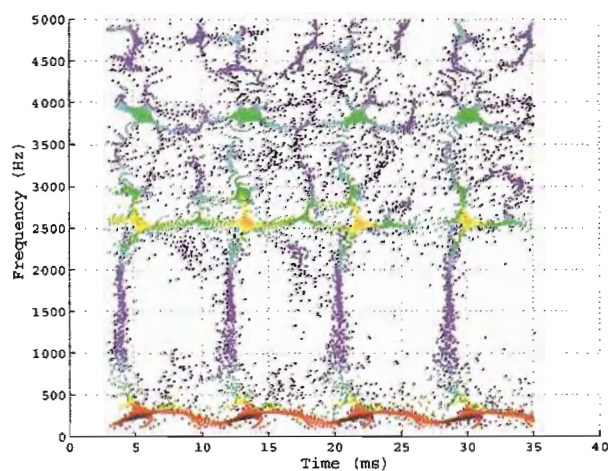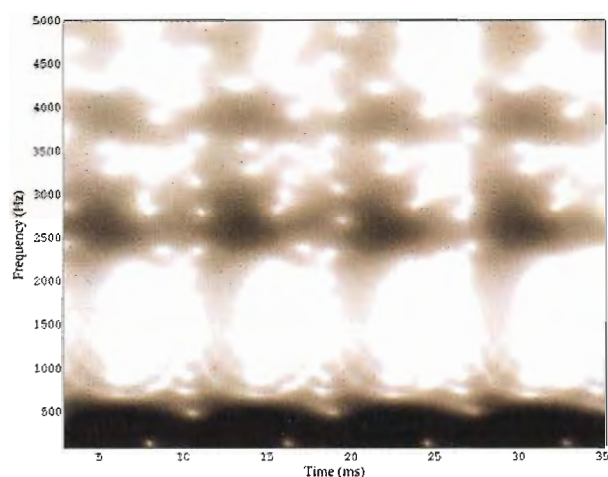

Fig. 5. Conventional and reassigned spectrograms of the same vowel as in Fig. 4, this time pronounced normally within an English word; computed using 5.9 ms frames. Note how the pattern of vocal tract excitations has changed from Fig. 4, although the same vowel is pronounced by the same speaker. Differences are largely due to the different voice quality of normal phonation.

A signal resulting from a pair of brown ghost fish (*Apteronotus leptorhynchus*) interacting can display at least two such components, and the fish have been described as abruptly moving the frequencies of their signals quite considerably in response to one another. Recent analysis performed by the first author in collaboration with John Lewis and Ginette Hupe of the University of Ottawa, however, reveals that these interactive modulations may instead involve only a slight adjustment of signal frequencies to induce different beat frequencies in the resulting combination tone. The auditory impression of a brief change in beating is often that of a chirp or abrupt change in fundamental frequency. Without the precision of the reassigned spectrogram, the resolution and measurement of closely aligned and beating gymnotid signals has been a serious challenge in the past. Figure 6 shows both a long and a short frame analysis of the same brown ghost signals; the long frame resolves the closely spaced line components, but does not have sufficient time resolution to show the beating between them, while the short frame fails to resolve the multiple components, but gains the ability to show the individual beats as impulse-like signal elements.
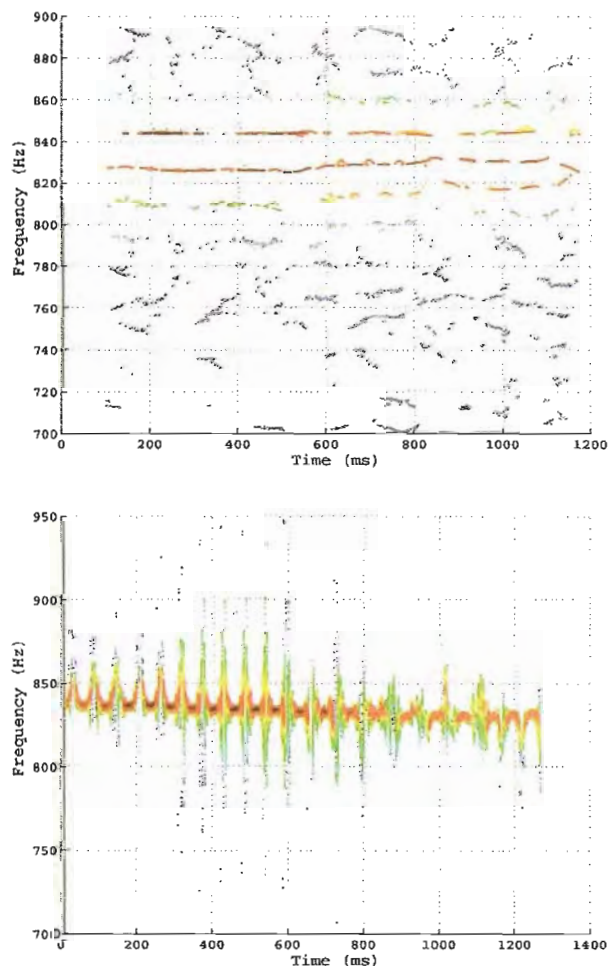

Fig. 6. Brown ghost fish signals interacting; upper panel analyzed with 200 ms frames, lower panel with 10 ms frames. The long frame reveals the small changes in fundamental frequencies of closely spaced components, while the short frame is unable to resolve the multiple components but shows the beats as quasi-impulsive amplitude modulations.

Sound modeling

Additive sound models represent sounds as a collection of amplitude- and frequency-modulated sinusoids. The time-varying frequencies and amplitudes are estimated from peaks in the spectrogram. These models have the very desirable property of easy and intuitive manipulability. Their parameters are easy to understand and deformations of the model data yield predictable results. Unfortunately, for many kinds of sounds, it is extremely difficult, using conventional techniques, to obtain a robust sinusoidal model that preserves all relevant characteristics of the original sound without introducing artifacts.

For example, an acoustic bass pluck is difficult to model because it requires very high temporal resolution to represent the abrupt attack without smearing. In fact, in order to capture the transient behavior of the pluck using a conventional spectrogram, a window much shorter than a single period of the waveform (which is approximately 13.6 ms in the example shown in Fig. 7) is needed. Any window that achieves the desired temporal resolution in a conventional approach will fail to resolve the harmonic components.

An additive model constructed by following ridges on a reassigned spectrogram yields greater precision in time and frequency than is possible using conventional additive techniques.[15] From the reassigned spectral data shown in Fig. 7, we can construct a robust model of the bass pluck that captures both the harmonic components in the decay and the transient behavior of the abrupt attack. Sounds reconstructed from a reassigned additive model preserve the temporal envelope of the original signal. Time-warping, pitch shifting, and sound morphing operations can all be performed on the model data while retaining the character of the original sound.

It is useful to emphasize that the single attack transient in the bass pluck has been located in time to a high level of precision using long analysis windows which resolve the harmonics, a feat that is not possible with the conventional spectrogram. This is not to suggest that two closely spaced transients could be resolved if they both fell within a single analysis window.

Improving on the improved

In spite of the obvious gains in clarity of the locations and movements of line components in these reassigned spectrograms, as well as the improved time localization of impulsive events, the images can be cluttered with meaningless random points. This is mainly because the algorithm employed to locate the AM/FM components in the signal has a meaningful output only in the neighborhood of a component. Where there is no component of significant amplitude, the time-frequency locations of the points to be plotted can become random.

We next describe a technique theoretically outlined by Doug Nelson of the National Security Administration at Fort Meade, which has the potential to "denoise" our spectrograms, and also to permit quasistationary (low FM rate) components to be isolated in a display, or alternatively to permit highly time-localized points (impulses) to be isolated.

Fig. 8. Here the reassigned spectrogram of Fig. 4 has been plotted to show only those points which are part of a nearly stationary line component, found using Nelson's higher-order partial derivative condition. These points have a frequency derivative of the instantaneous frequency within 0.25 of 0.
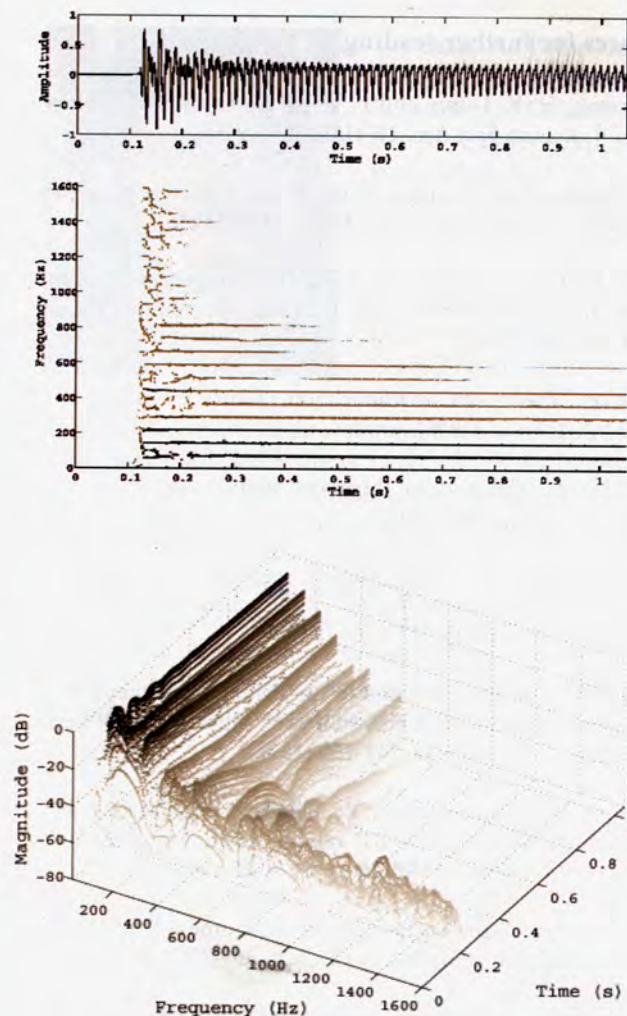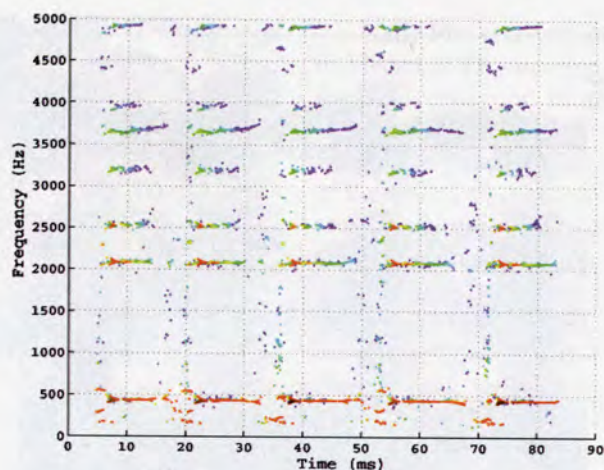
Fig. 7. Reassigned spectrograms of 70 Hz acoustic bass pluck, computed using a Kaiser window of 1901 samples at 44.1 kHz. Both the transient attack and the harmonics are able to be captured with one choice of analysis frame. Lower panel shows a different view of the same 3D scatterplot, making it clear how the ridges could be tracked for additive sound modeling.

Nelson[16] explained that the nearly stationary AM/FM components of a signal have a frequency derivative of the instantaneous frequency near zero. By plotting just those points in a reassigned spectrogram meeting this condition on the higher-order mixed partial STFT phase derivative to within a threshold, a spectrogram showing just the line components can be drawn (Fig. 8). The precise threshold can be empirically determined, and will in practice depend on the degree of deviation from a pure sinusoid that is tolerable in the application at hand.

Nelson further asserted the dual fact that the impulses in a signal have a value for this same mixed partial phase derivative near 1. By plotting just those points meeting this condition to within a threshold, a spectrogram showing just the impulsive events in a signal can alternatively be drawn (Fig. 9). Plotting all points meeting the disjunction of the above conditions results in a denoised spectrogram showing quasi-sinusoidal components and impulses together, to the exclusion of most everything else. Figure 10 shows another demonstration of the usefulness of Nelson's idea; the excited

vocal tract resonance frequencies are much easier to measure in the lower panel, where they are isolated, than in the upper spectrogram.

Summary

The reassigned spectrogram is a technique for analyzing sound and other signals into their AM/FM components, and showing the time course of the amplitudes and instantaneous frequencies of these components in a 3D plot. While similar in spirit to the 60-year old spectrogram which it improves upon, it embodies a fundamental departure from the time-worn efforts to show the complete energy distribution of a signal in time and frequency. In working to perfect the method as well as our understanding of the resulting images, the researchers who have contributed to this technology over the past 30 years—sometimes in ignorance of
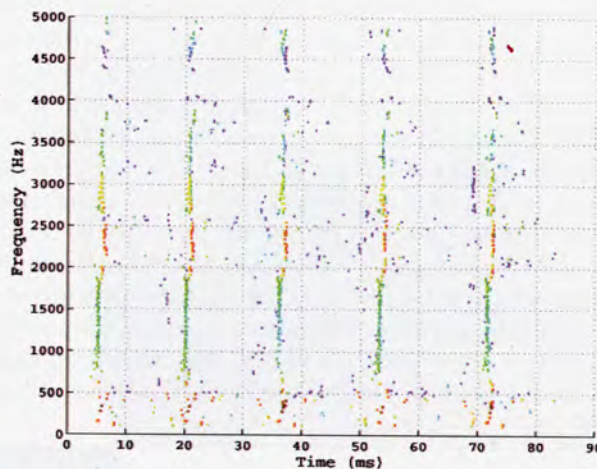

Fig. 9. Here the reassigned spectrogram of Fig. 4 has been plotted to show only those points which are part of an impulse, found using Nelson's higher-order partial derivative condition. These points have a value for the derivative within 0.25 of 1.
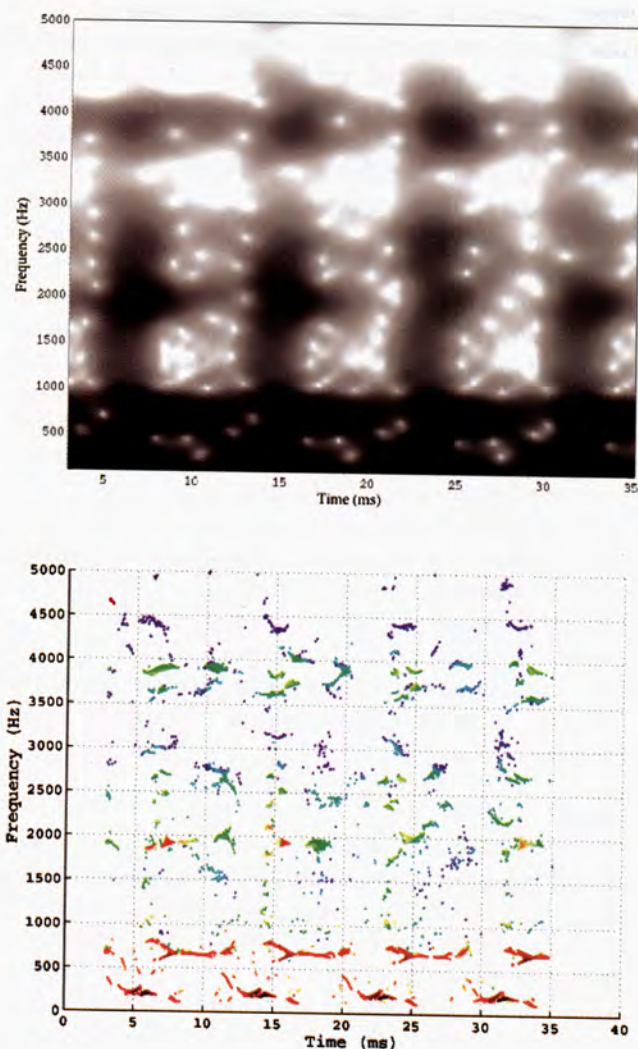
Fig. 10. A few milliseconds from vowel [ae] "had," comparing the visibility of resonance components in a conventional spectrogram (upper) against a reassigned spectrogram plotted using Nelson's component isolation technique. Both are computed using 5.9 ms analysis frames.

References for further reading:

1 W. Koenig, H. K. Dunn, and L. Y. Lacy, "The sound spectrograph," J. Acoust. Soc. Am. **18**(1), 19-49 (1946).

2 G. A. Kopp and H. C. Green, "Basic phonetic principles of visible speech," J. Acoust. Soc. Am. **18**(1), 74-89 (1946).

3 L. K. Montgomery and I. S. Reed, "A generalization of the Gabor-Helstrom transform," IEEE Trans. Information Theory **IT-13**, 344-345 (1967).

4 D. Gabor, "Theory of Communication," J. Inst. Elect. Eng. Part III **93**(26), 429-457 (1946).

5 J. R. Carson, "Notes of the theory of modulation," Proc. Inst. Radio Eng. **10**(1), 57-64 (1922).

6 J. R. Carson and T. C. Fry, "Variable frequency electric circuit theory with application to the theory of frequency-modulation," Bell Syst. Tech. J. **16**(4), 513-540 (1937).

7 K. Kodera, C. de Villedary, and R. Gendrin, "A new method for the numerical analysis of non-stationary signals," Phys. Earth and Planetary Interiors **12**, 142-150 (1976).

8 K. Kodera, R. Gendrin, and C. de Villedary, "Analysis of time-varying signals with small BT values," IEEE Trans. Acoust., Speech, and Sig. Proc. **ASSP-26**(1), 64-76 (1978).

9 A. W. Rihaczek, "Signal energy distribution in time and frequency," IEEE Trans. Information Theory **IT-14**(3), 369-374 (1968).

10 D. J. Nelson, "Cross-spectral methods for processing speech," J. Acoust. Soc. Am. **110**(5), 2575-2592 (2001).

11 T. Nakatani and T. Irino, "Robust and accurate fundamental frequency estimation based on dominant harmonic components," J. Acoust. Soc. Am. **116**(6), 3690-3700 (2004).

12 F. Auger and P. Flandrin, "Improving the readability of time-frequency and time-scale representations by the reassignment method," IEEE Trans. Sig. Proc. **43**(5), 1068-1089 (1995).

13 S. A. Fulop and K. Fitz, "Algorithms for computing the time-corrected instantaneous frequency (reassigned) spectrogram, with applications," J. Acoust. Soc. Am. **119**(1), 360-371 (2006).

14 J. L. Larimer and J. A. MacDonald, "Sensory feedback from electroreceptors to electromotor pacemaker centers in gymnotids," Am. J. Physiol. **214**(6), 1253-1261 (1968).

15 K. Fitz and L. Haken, "On the use of time-frequency reassignment in additive sound modeling," J. Audio Eng. Soc. **50**(11), 879-893 (2002).

16 D. J. Nelson, "Instantaneous higher order phase derivatives," Digital Sig. Proc. **12**, 416-428 (2002).

each other's work—have each recognized that, for many applications, it is the components of a signal that contain the most sought information, not the overall energy distribution. Going hand in hand with this break from tradition, by focusing entirely on instantaneous frequencies the method sets aside the "classical" Fourier spectrum which decomposes a signal (or its successive frames, in the spectrogram) into sine waves of infinite extent. From the perspective of signal modeling, we already know that real signals are never so constituted, so it is not exactly a leap of faith to stop representing them that way. It is our hope that the future will bring production-quality software that everyone can access, which allows a choice between conventional and reassigned spectrograms. For many applied acousticians, the choice will be clear.**AT**

Sean A. Fulop received a B.Sc. in Physics (1991) from the University of Calgary and, after two M.A. degrees in Linguistics, received his Ph.D. in Linguistics (1999) from the University of California, Los Angeles. He then held temporary faculty positions in Linguistics at San José State University and the University of Chicago before joining the faculty at California State University, Fresno, as an Assistant Professor of Linguistics in 2005. His publication areas range widely in linguistics and speech and cover speech processing, phonetics, mathematical linguistic theory, and computational linguistics. His chief research programs in acoustics involve the investigation of speech sounds, and the development, evaluation, and dissemination of improved signal processing tools for phonetics and speech acoustics research. He currently reviews speech-related patents for the *Journal of the Acoustical Society of America*, and has been a member of the Society since 1987.

Kelly Fitz received a Ph.D. in Electrical Engineering from the University of Illinois at Urbana-Champaign in 1999. He was the principal developer of the sound analysis and synthesis software Lemur and at the National Center for Supercomputing Applications. He was co-developer of the Vanilla Sound Server, an application providing real-time sound synthesis for virtual environments. Dr. Fitz is currently Assistant Professor of Electrical Engineering and Computer Science at Washington State University, and the principal developer of the Loris software for sound modeling and morphing. His research interests include speech and audio processing, digital sound synthesis, and computer music composition.