On the Use of Time–Frequency Reassignment in Additive Sound Modeling*

KELLY FITZ, *AES Member* **AND LIPPOLD HAKEN,** *AES Member*

Department of Electrical Engineering and Computer Science, Washington University, Pulman, WA 99164

A method of reassignment in sound modeling to produce a sharper, more robust additive representation is introduced. The reassigned bandwidth-enhanced additive model follows ridges in a time–frequency analysis to construct partials having both sinusoidal and noise characteristics. This model yields greater resolution in time and frequency than is possible using conventional additive techniques, and better preserves the temporal envelope of transient signals, even in modified reconstruction, without introducing new component types or cumbersome phase interpolation algorithms.

0 INTRODUCTION

The method of reassignment has been used to sharpen spectrograms in order to make them more readable [1], [2], to measure sinusoidality, and to ensure optimal window alignment in the analysis of musical signals [3]. We use time–frequency reassignment to improve our bandwidth-enhanced additive sound model. The bandwidth-enhanced additive representation is in some way similar to traditional sinusoidal models [4]–[6] in that a waveform is modeled as a collection of components, called partials, having time-varying amplitude and frequency envelopes. Our partials are not strictly sinusoidal, however. We employ a technique of bandwidth enhancement to combine sinusoidal energy and noise energy into a single partial having time-varying amplitude, frequency, and bandwidth parameters [7], [8].

Additive sound models applicable to polyphonic and nonharmonic sounds employ long analysis windows, which can compromise the time resolution and phase accuracy needed to preserve the temporal shape of transients. Various methods have been proposed for representing transient waveforms in additive sound models. Verma and Meng [9] introduce new component types specifically for modeling transients, but this method sacrifices the homogeneity of the model. A homogeneous model, that is, a model having a single component type, such as the breakpoint parameter envelopes in our reassigned bandwidth-enhanced additive model [10], is critical for many kinds of manipulations [11], [12]. Peeters and Rodet [3] have developed a hybrid analysis/synthesis system that eschews high-level transient models and retains unabridged OLA (overlap–add) frame data at transient positions. This hybrid representation represents unmodified transients perfectly, but also sacrifices homogeneity. Quatieri et al. [13] propose a method for preserving the temporal envelope of short-duration complex acoustic signals using a homogeneous sinusoidal model, but it is inapplicable to sounds of longer duration, or sounds having multiple transient events.

We use the method of reassignment to improve the time and frequency estimates used to define our partial parameter envelopes, thereby enhancing the time–frequency resolution of our representation, and improving its phase accuracy. The combination of time–frequency reassignment and bandwidth enhancement yields a homogeneous model (that is, a model having a single component type) that is capable of representing at high fidelity a wide variety of sounds, including nonharmonic, polyphonic, impulsive, and noisy sounds. The reassigned bandwidth-enhanced sound model is robust under transformation, and the fidelity of the representation is preserved even under time dilation and other model-domain modifications. The homogeneity and robustness of the reassigned bandwidth-enhanced model make it particularly well suited for such manipulations as cross synthesis and sound morphing.

Reassigned bandwidth-enhanced modeling and rendering and many kinds of manipulations, including morphing, have been implemented in the open-source C++ class library Loris [14], and a stream-based, real-time implementation of bandwidth-enhanced synthesis is available in the Symbolic Sound Kyma environment [15].

* Manuscript received 2001 December 20; revised 2002 July 23 and 2002 September 11.

1 TIME–FREQUENCY REASSIGNMENT

The discrete short-time Fourier transform is often used as the basis for a time–frequency representation of time-varying signals, and is defined as a function of time index n and frequency index k as

$$X_n(k) = \sum_{l=-\infty}^{\infty} h(l-n)\,x(l)\,\exp\left[\frac{-\mathrm{j}2\pi(l-n)k}{N}\right] \quad (1)$$

$$= \sum_{l=-\frac{N-1}{2}}^{\frac{N-1}{2}} h(l)\,x(n+l)\,\exp\left(\frac{-\mathrm{j}2\pi lk}{N}\right) \quad (2)$$

where $h(n)$ is a sliding window function equal to 0 for $n < -(N-1)/2$ and $n > (N-1)/2$ (for N odd), so that $X_n(k)$ is the N-point discrete Fourier transform of a short-time waveform centered at time n.

Short-time Fourier transform data are sampled at a rate equal to the analysis hop size, so data in derivative time–frequency representations are reported on a regular temporal grid, corresponding to the centers of the short-time analysis windows. The sampling of these so-called frame-based representations can be made as dense as desired by an appropriate choice of hop size. However, temporal smearing due to long analysis windows needed to achieve high-frequency resolution cannot be relieved by denser sampling.

Though the short-time phase spectrum is known to contain important temporal information, typically only the short-time magnitude spectrum is considered in the time–frequency representation. The short-time phase spectrum is sometimes used to improve the frequency estimates in the time–frequency representation of quasi-harmonic sounds [16], but it is often omitted entirely, or used only in unmodified reconstruction, as in the basic sinusoidal model, described by McAulay and Quatieri [4].

The so-called method of reassignment computes sharpened time and frequency estimates for each spectral component from partial derivatives of the short-time phase spectrum. Instead of locating time–frequency components at the geometrical center of the analysis window (t_n, ω_k), as in traditional short-time spectral analysis, the components are reassigned to the center of gravity of their complex spectral energy distribution, computed from the short-time phase spectrum according to the principle of stationary phase [17, ch. 7.3]. This method was first developed in the context of the spectrogram and called the modified moving window method [18], but it has since been applied to a variety of time–frequency and time-scale transforms [1].

The principle of stationary phase states that the variation of the Fourier phase spectrum not attributable to periodic oscillation is slow with respect to frequency in certain spectral regions, and in surrounding regions the variation is relatively rapid. In Fourier reconstruction, positive and negative contributions to the waveform cancel in frequency regions of rapid phase variation. Only regions of slow phase variation (stationary phase) will contribute significantly to the reconstruction, and the maximum contribution (center of gravity) occurs at the point where the phase is changing most slowly with respect to time and frequency.

In the vicinity of $t = \tau$ (that is, for an analysis window centered at time $t = \tau$), the point of maximum spectral energy contribution has time–frequency coordinates that satisfy the stationarity conditions

$$\frac{\partial}{\partial\omega}\left[\phi(\tau,\omega) + \omega(t-\tau)\right] = 0 \quad (3)$$

$$\frac{\partial}{\partial\tau}\left[\phi(\tau,\omega) + \omega(t-\tau)\right] = 0 \quad (4)$$

where $\phi(\tau, \omega)$ is the continuous short-time phase spectrum and $\omega(t-\tau)$ is the phase travel due to periodic oscillation [18]. The stationarity conditions are satisfied at the coordinates

$$\hat{t} = \tau - \frac{\partial\phi(\tau,\omega)}{\partial\omega} \quad (5)$$

$$\hat{\omega} = \frac{\partial\phi(\tau,\omega)}{\partial\tau} \quad (6)$$

representing group delay and instantaneous frequency, respectively.

Discretizing Eqs. (5) and (6) to compute the time and frequency coordinates numerically is difficult and unreliable, because the partial derivatives must be approximated. These formulas can be rewritten in the form of ratios of discrete Fourier transforms [1]. Time and frequency coordinates can be computed using two additional short-time Fourier transforms, one employing a time-weighted window function and one a frequency-weighted window function.

Since time estimates correspond to the temporal center of the short-time analysis window, the time-weighted window is computed by scaling the analysis window function by a time ramp from $-(N-1)/2$ to $(N-1)/2$ for a window of length N. The frequency-weighted window is computed by wrapping the Fourier transform of the analysis window to the frequency range $[-\pi, \pi]$, scaling the transform by a frequency ramp from $-(N-1)/2$ to $(N-1)/2$, and inverting the scaled transform to obtain a (real) frequency-scaled window. Using these weighted windows, the method of reassignment computes corrections to the time and frequency estimates in fractional sample units between $-(N-1)/2$ to $(N-1)/2$. The three analysis windows employed in reassigned short-time Fourier analysis are shown in Fig. 1.

The reassigned time $\hat{t}_{k,n}$ for the kth spectral component from the short-time analysis window centered at time n (in samples, assuming odd-length analysis windows) is [1]

$$\hat{t}_{k,n} = n - \Re\left[\frac{X_{t;n}(k)\,X_n^*(k)}{\left|X_n(k)\right|^2}\right] \quad (7)$$

where $X_{t;n}(k)$ denotes the short-time transform computed using the time-weighted window function and $\Re\,[\cdot]$ denotes the real part of the bracketed ratio.

The corrected frequency $\hat{\omega}_{k,n}(k)$ corresponding to the same component is [1]

$$\hat{\omega}_{k,n} = k + \Im\left[\frac{X_{f;n}(k)\,X_n^*(k)}{\left|X_n(k)\right|^2}\right] \tag{8}$$

where $X_{f;n}(k)$ denotes the short-time transform computed using the frequency-weighted window function and $\Im[\cdot]$ denotes the imaginary part of the bracketed ratio. Both $t_{k,n}$ and $\hat{\omega}_{k,n}$ have units of fractional samples.

Time and frequency shifts are preserved in the reassignment operation, and energy is conserved in the reassigned time–frequency data. Moreover, chirps and impulses are perfectly localized in time and frequency in any reassigned time–frequency or time-scale representation [1]. Reassignment sacrifices the bilinearity of time–frequency transformations such as the squared magnitude of the short-time Fourier transform, since very data point in the representation is relocated by a process that is highly signal dependent. This is not an issue in our representation, since the bandwidth-enhanced additive model, like the basic sinusoidal model [4], retains data only at time–frequency ridges (peaks in the short-time magnitude spectra), and thus is not bilinear.

Note that since the short-time Fourier transform is invertible, and the original waveform can be exactly reconstructed from an adequately sampled short-time Fourier representation, all the information needed to precisely locate a spectral component within an analysis window is present in the short-time coefficients $X_n(k)$. Temporal information is encoded in the short-time phase spectrum, which is very difficult to interpret. The method reassignment is a technique for extracting information from the phase spectrum.

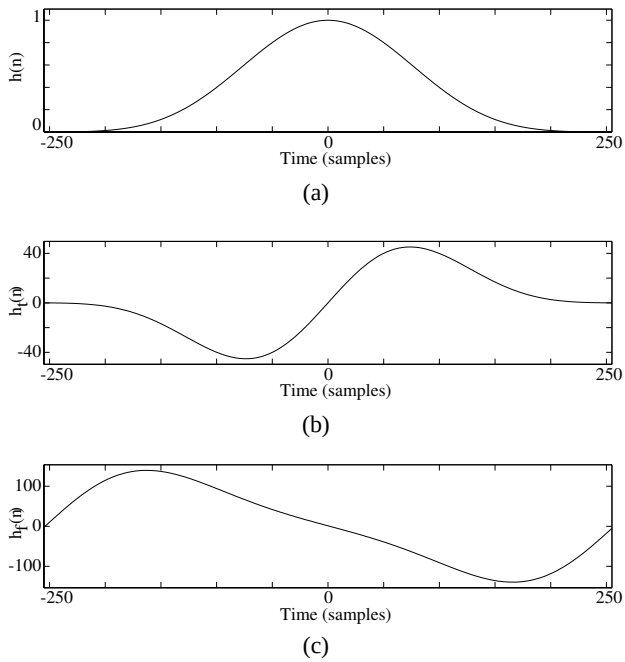


Fig. 1. Analysis windows employed in three short-time transforms used to compute reassigned times and frequencies. (a) Original window function $h(n)$ (a 501-point Kaiser window with shaping parameter 12.0 in this case). (b) Time-weighted window function $h_t(n) = nh(n)$. (c) Frequency-weighted window function $h_f(n)$.

2 REASSIGNED BANDWIDTH-ENHANCED ANALYSIS

The reassigned bandwidth-enhanced additive model [10] employs time–frequency reassignment to improve the time and frequency estimates used to define partial parameter envelopes, thereby improving the time–frequency resolution and the phase accuracy of the representation. Reassignment transforms our analysis from a frame-based analysis into a "true" time–frequency analysis. Whereas the discrete short-time Fourier transform defined by Eq. (2) orients data according to the analysis frame rate and the length of the transform, the time and frequency orientation of reassigned spectral data is solely a function of the data themselves.

The method of analysis we use in our research models a sampled audio waveform as a collection of bandwidth-enhanced partials having sinusoidal and noiselike characteristics. Other methods for capturing noise in additive sound models [5], [19] have represented noise energy in fixed frequency bands using more than one component type. By contrast, bandwidth-enhanced partials are defined by a trio of synchronized breakpoint envelopes specifying the time-varying amplitude, center frequency, and noise content for each component. Each partial is rendered by a bandwidth-enhanced oscillator, described by

$$y(n) = \left[A(n) + \mathrm{B}(n)\zeta(n)\right]\cos\left[\theta(n)\right] \tag{9}$$

where $A(n)$ and $\beta(n)$ are the time-varying sinusoidal and noise amplitudes, respectively, and $\zeta(n)$ is a energy-normalized low-pass noise sequence, generated by exciting a low-pass filter with white noise and scaling the filter gain such that the noise sequence has the same total spectral energy as a full-amplitude sinusoid. The oscillator phase $\theta(n)$ is initialized to some starting value, obtained from the reassigned short-time phase spectrum, and updated according to the time-varying radian frequency $\omega(n)$ by

$$\theta(n) = \theta(n-1) + \omega(n), \quad n > 0 \tag{10}$$

The bandwidth-enhanced oscillator is depicted in Fig. 2.

We define the time-varying bandwidth coefficient $\kappa(n)$ as the fraction of total instantaneous partial energy that is attributable to noise. This bandwidth (or noisiness) coefficient assumes values between 0 for a pure sinusoid and 1 for a partial that is entirely narrow-band noise, and varies over time according to the noisiness of the partial. If we represent the total (sinusoidal and noise) instantaneous partial energy as $\tilde{A}^2(n)$, then the output of the bandwidth-enhanced oscillator is described by

$$y(n) = \tilde{A}(n)\left[\sqrt{1-\kappa(n)} + \sqrt{2\kappa(n)}\,\zeta(n)\right]\cos\left[\theta(n)\right]. \tag{11}$$

The envelopes for the time-varying partial amplitudes and frequencies are constructed by identifying and following

the ridges on the time–frequency surface. The time-varying partial bandwidth coefficients are computed and assigned by a process of bandwidth association [7].

We use the method of reassignment to improve the time and frequency estimates for our partial parameter envelope breakpoints by computing reassigned times and frequencies that are not constrained to lie on the time–frequency grid defined by the short-time Fourier analysis parameters. Our algorithm shares with traditional sinusoidal methods the notion of temporally connected partial parameter estimates, but by contrast, our estimates are nonuniformly distributed in both time and frequency.

Short-time analysis windows normally overlap in both time and frequency, so time–frequency reassignment often yields time corrections greater than the length of the short-time hop size and frequency corrections greater than the width of a frequency bin. Large time corrections are common in analysis windows containing strong transients that are far from the temporal center of the window. Since we retain data only at time–frequency ridges, that is, at frequencies of spectral energy concentration, we generally observe large frequency corrections only in the presence of strong noise components, where phase stationarity is a weaker effect.

3 SHARPENING TRANSIENTS

Time–frequency representations based on traditional magnitude-only short-time Fourier analysis techniques (such as the spectrogram and the basic sinusoidal model [4]) fail to distinguish transient components from sustaining components. A strong transient waveform, as shown in Fig. 3(a), is represented by a collection of low-amplitude spectral components in early short-time analysis frames, that is, frames corresponding to analysis windows centered earlier than the time of the transient. A low-amplitude periodic waveform, as shown in Fig. 3(b), is also represented by a collection of low-amplitude spectral components. The information needed to distinguish these two critically different waveforms is encoded in the short-time phase spectrum, and is extracted by the method of reassignment.

Time–frequency reassignment allows us to preserve the temporal envelope shape without sacrificing the homogeneity of the bandwidth-enhanced additive model. Components extracted from early or late short-time analysis windows are relocated nearer to the times of transient events, yielding clusters of time–frequency data points, as depicted in Fig. 4. In this way, time reassignment greatly reduces the temporal smearing introduced through the use of long analysis windows. Moreover, since reassignment sharpens our frequency estimates, it is possible to achieve good frequency resolution with shorter (in time) analysis windows than would be possible with traditional methods. The use of shorter analysis windows further improves our time resolution and reduces temporal smearing.

The effect of time–frequency reassignment on the transient response can be demonstrated using a square wave that turns on abruptly, such as the waveform shown in Fig. 5. This waveform, while aurally uninteresting and uninformative, is useful for visualizing the performance of various analysis methods. Its abrupt onset makes temporal smearing obvious, its simple harmonic partial amplitude relationship makes it easy to predict the necessary data for a good time–frequency representation, and its simple waveshape makes phase errors and temporal distortion easy to identify. Note, however, that this waveform is pathological for Fourier-based additive models, and exaggerates all of these problems with such methods. We use it only for the comparison of various methods.

Fig. 6 shows two reconstructions of the onset of a square wave from time–frequency data obtained using overlapping 54-ms analysis windows, with temporal centers separated by 10 ms. This analysis window is long compared to the period of the square wave, but realistic for the case of a polyphonic sound (a sound having multiple simultaneous voices), in which the square wave is one voice. For clarity, only the square wave is presented in this example, and other simultaneous voices are omitted. The square wave

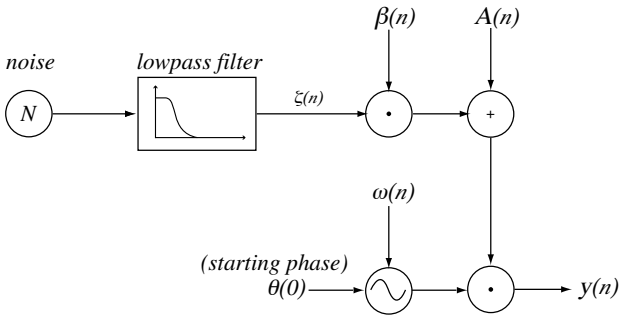
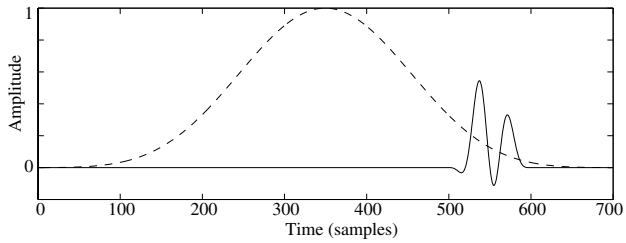
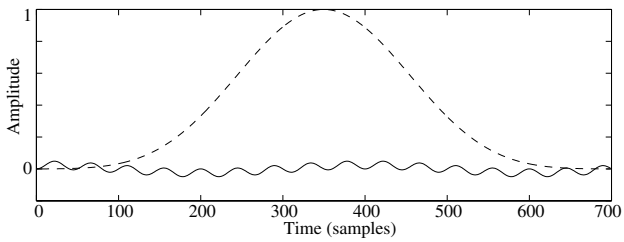


Fig. 2. Block diagram of bandwidth-enhanced oscillator. Time-varying sinusoidal and noise amplitudes are controlled by $A(n)$ and $\beta(n)$, respectively; time-varying center (sinusoidal) frequency is $\omega(n)$.



(a)



(b)

Fig. 3. Windowed short-time waveforms (dashed lines), not readily distinguished in basic sinsoidal model [4]. Both waveforms are represented by low-amplitude spectral components. (a) Strong transient yields off-center components, having large time corrections (positive in this case because transient is near right tail of window). (b) Sustained quasi-periodic waveform yields time corrections near zero.

has an abrupt onset. The silence before the onset is not shown. Only the first (lowest frequency) five harmonic partials were used in the reconstruction, and consequently the ringing due to Gibb's phenomenon is evident.

Fig. 6(a) is a reconstruction from traditional, nonreassigned time–frequency data. The reconstructed square wave amplitude rises very gradually and reaches full amplitude approximately 40 ms after the first nonzero sample. Clearly, the instantaneous turn-on has been smeared out by the long analysis window. Fig. 6(b) shows a reconstruction from reassigned time–frequency data. The transient response has been greatly improved by relocating components extracted from early analysis windows (like the one on the left in Fig. 5) to their spectral centers of gravity, closer to the observed turn-on time. The synthesized onset time has been reduced to approximately 10 ms. The corresponding time–frequency analysis data are shown in Fig. 7. The nonreassigned data are evenly distributed in time, so data from early windows (that is, windows centered before the onset time) smear the onset, whereas the reassigned data from early analysis windows are clumped near the correct onset time.

4 CROPPING

Off-center components are short-time spectral components having large time reassignments. Since they represent transient events that are far from the center of the analysis window, and are therefore poorly represented in the windowed short-time waveform, these off-center components introduce unreliable spectral parameter estimates that corrupt our representation, making the model data difficult to interpret and manipulated.

Fortunately large time corrections make off-center components easy to identify and remove from our model. By removing the unreliable data embodied by off-center components, we make our model cleaner and more robust. Moreover, thanks to the redundancy inherent in short-time analysis with overlapping analysis windows, we do not sacrifice information by removing the unreliable data points. The information represented poorly in off-center components is more reliably represented in well-centered components, extracted from analysis windows centered nearer the time of the transient event. Typically, data having time corrections

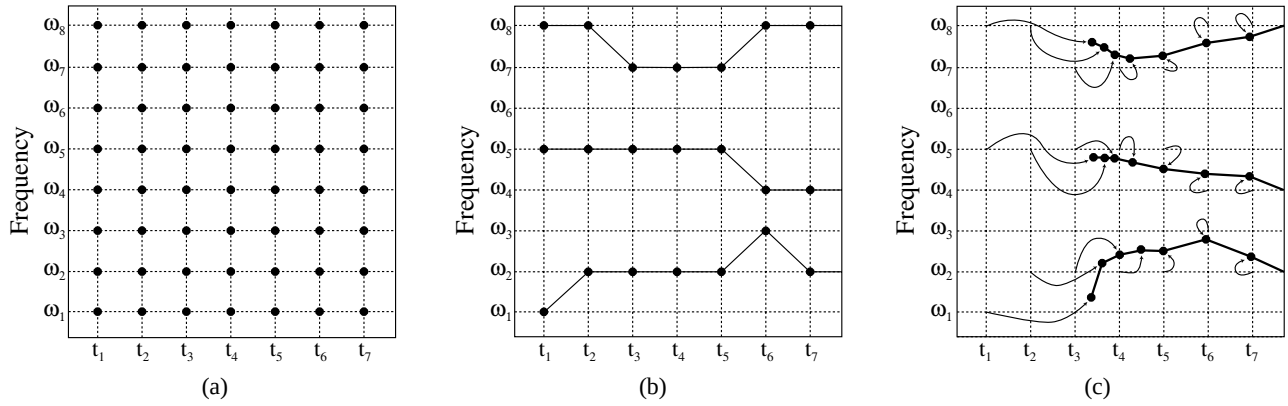


Fig. 4. Comparison of time–frequency data included in common representations. Only time–frequency orientation of data points is shown. (a) Short-time Fourier transform retains data at every time t_n and frequency ω_k. (b) Basic sinusoidal model [4] retains data at selected time and frequency samples. (c) Reassigned bandwidth-enhanced analysis data are distributed continuously in time and frequency, and retained only at time–frequency ridges. Arrows indicate mapping of short-time spectral samples onto time–frequency ridges due to method of reassignment.

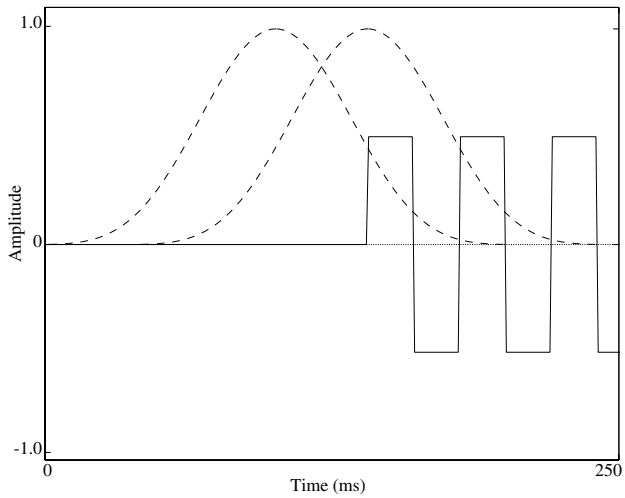


Fig. 5. Two long analysis windows superimposed at different times on square wave signal with abrupt turn-on. Short-time transform corresponding to earlier window generates unreliable parameter estimates and smears sharp onset of square wave.

greater than the time between consecutive analysis window centers are considered to be unreliable and are removed, or cropped.

Cropping partials to remove off-center components allows us to localize transient events reliably. Fig. 7(c) shows reassigned time–frequency data from the abrupt square wave onset with off-center components removed. The abrupt square wave onset synthesized from the cropped reassigned data, seen in Fig. 6(c), is much sharper than the uncropped reassigned reconstruction, because the taper of the analysis window makes even the time correction data unreliable in components that are very far off center.

Fig. 8 shows reassigned bandwidth-enhanced model data from the onset of a bowed cello tone before and after the removal of off-center components. In this case, components with time corrections greater than 10 ms (the time between consecutive analysis windows) were deemed to be too far off center to deliver reliable parameter estimates. As in Fig. 7(c), the unreliable data clustered at the time of the onset are removed, leaving a cleaner, more robust representation.

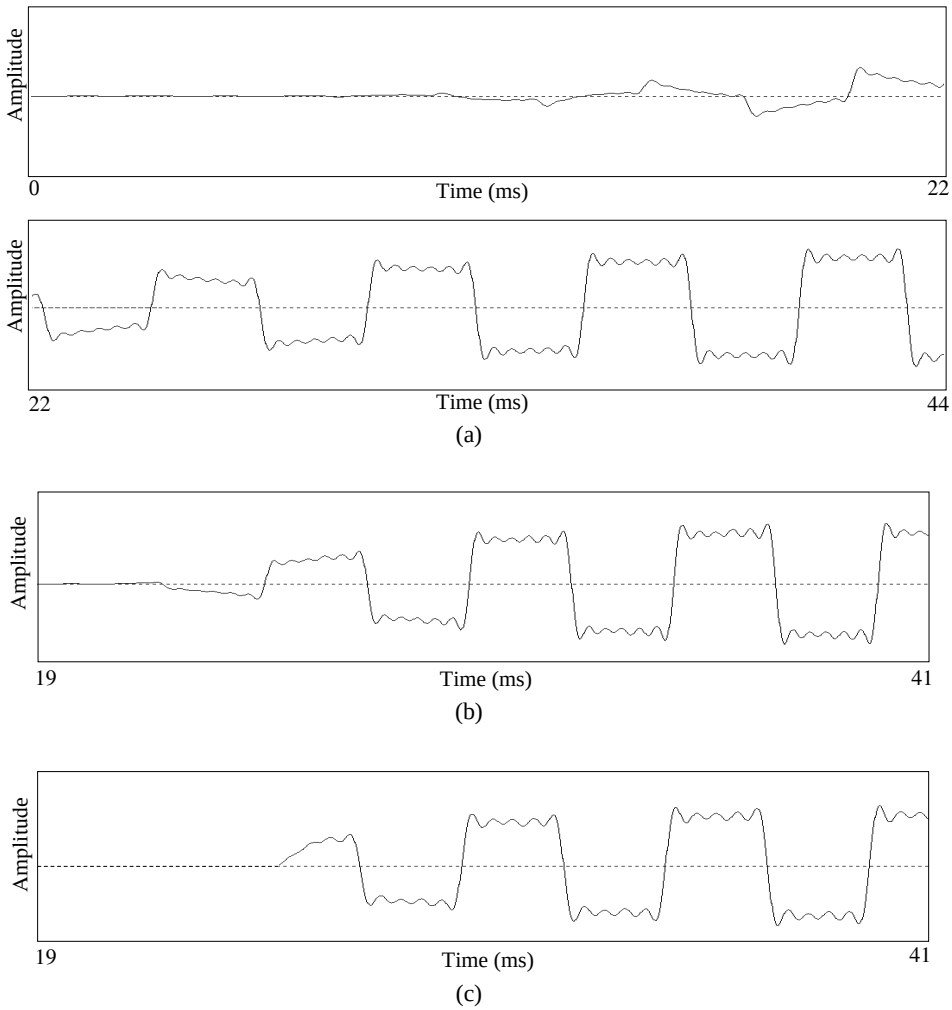


Fig. 6. Abrupt square wave onset reconstructed from five sinusoidal partials corresponding to first five harmonics. (a) Reconstruction from nonreassigned analysis data. (b) Reconstruction from reassigned analysis data. (c) Reconstruction from reassigned analysis data with unreliable partial parameter estimates removed, or cropped.

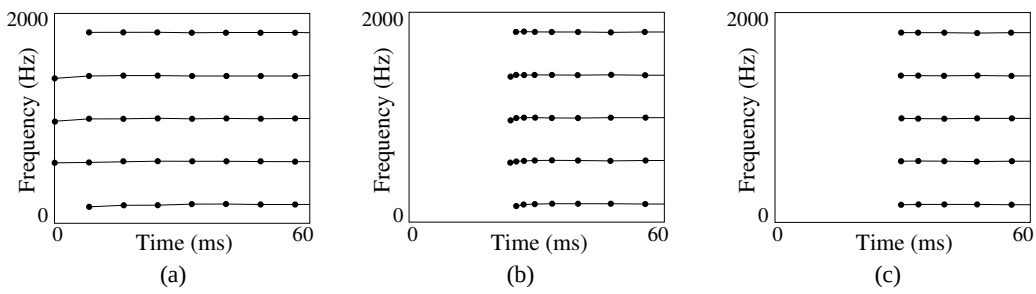


Fig. 7. Time–frequency analysis data points for abrupt square wave onset. (a) Traditional nonreassigned data are evenly distributed in time. (b) Reassigned data are clumped at onset time. (c) Reassigned analysis data after far off-center components have been removed, or cropped. Only time and frequency information is plotted; amplitude information is not displayed.

5 PHASE MAINTENANCE

Preserving phase is important for reproducing some classes of sounds, in particular transients and short-duration complex audio events having significant information in the temporal envelope [13]. The basic sinusoidal models proposed by McAulay and Quatieri [4] is phase correct, that is, it preserves phase at all times in unmodified reconstruction. In order to match short-time spectral frequency and phase estimates at frame boundaries, McAulay and Quatieri employ cubic interpolation of the instantaneous partial phase.

Cubic phase envelopes have many undesirable properties. They are difficult to manipulate and maintain under time- and frequency-scale transformation compared to linear frequency envelopes. However, in unmodified reconstruction, cubic interpolation prevents the propagation of phase errors introduced by unreliable parameter estimates, maintaining phase accuracy in transients, where the temporal envelope is important, and throughout the reconstructed waveform. The effect of phase errors in the unmodified reconstruction of a square wave is illustrated in Fig. 9. If not corrected using a technique such as cubic phase interpolation, partial parameter errors introduced by off-center components render the waveshape visually unrecognizable. Fig. 9(b) shows that cubic phase can be used to correct these errors in unmodified reconstruction.

It should be noted that, in this particular case, the phase errors appear dramatic, but do not affect the sound of the reconstructed steady-state waveforms appreciably. In many sounds, particularly transient sounds, preservation of the temporal envelope is critical [13], [9], but since they lack audible onset transients, the square waves in Fig. 9(a)–(c) sound identical. It should also be noted that cubic phase interpolation can be used to preserve phase accuracy, but does not reduce temporal smearing due to off-center components in long analysis windows.

It is not desirable to preserve phase at all times in modified reconstruction. Because frequency is the time derivative of phase, any change in the time or frequency scale of a partial must correspond to a change in the phase values at the parameter envelope breakpoints. In general, preserving phase using the cubic phase method in the presence of modifications (or estimation errors) introduces wild frequency excursions [20]. Phase can be preserved at one time, however, and that time is typically chosen to be the onset of each partial, although any single time could be chosen. The partial phase at all other times is modified to reflect the new time–frequency characteristic of the modified partial.

Off-center components with unreliable parameter estimates introduce phase errors in modified reconstruction. If the phase is maintained at the partial onset, even the cubic interpolation scheme cannot prevent phase errors from propagating in modified syntheses. This effect is illustrated in Fig. 9(c), in which the square wave time–frequency data have been shifted in frequency by 10% and reconstructed using cubic phase curves modified to reflect the frequency shift.

By removing the off-center components at the onset of a partial, we not only remove the primary source of phase errors, we also improve the shape of the temporal envelope in the modified reconstruction of transients by preserving a more reliable phase estimate at a time closer to the time of the transient event. We can therefore maintain phase accuracy at critical parts of the audio waveform even under transformation, and even using linear frequency envelopes, which are much simpler to compute, interpret, edit, and maintain than cubic phase curves. Fig. 9(d) shows a square wave reconstruction from cropped reassigned time–frequency data, and Fig. 9(e) shows a frequency-shifted reconstruction, both using linear frequency interpolation. Removing components with large time corrections preserves phase in modified and unmodified reconstruction, and thus obviates cubic phase interpolation.

Moreover, since we do not rely on frequent cubic phase corrections to our frequency estimates to preserve the shape of the temporal envelope (which would otherwise be corrupted by errors introduced by unreliable data), we have found that we can obtain very good-quality reconstruction, even under modification, with regularly sampled partial parameter envelopes. That is, we can sample the frequency, amplitude, and bandwidth envelopes of our

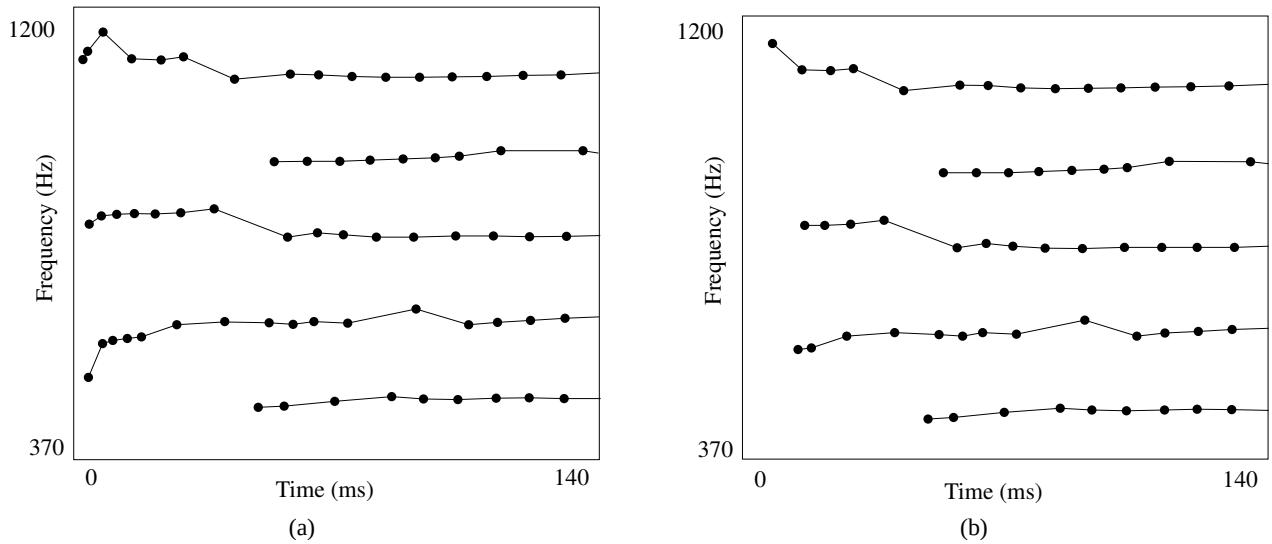


Fig. 8. Time–frequency coordinates of data from reassigned bandwidth-enhanced analysis. (a) Before cropping. (b) After cropping of off-center components clumped together at partial onsets. Source waveform is a bowed cello tone.

reassigned bandwidth-enhanced partials at regular intervals (of, for example, 10 ms) without sacrificing the fidelity of the model. We thereby achieve the data regularity of frame-based additive model data and the fidelity of reassigned spectral data. Resampling of the partial parameter envelopes is especially useful in real-time synthesis applications [11], [12].

6 BREAKING PARTIALS AT TRANSIENT EVENTS

Transients corresponding to the onset of all associated partials are preserved in our model by removing off-center components at the ends of partials. If transients always correspond to the onset of associated partials, then that method will preserve the temporal envelope of multiple transient events. In fact, however, partials often span transients. Fig. 10 shows a partial that extends over transient boundaries in a representation of a bongo roll, a sequence of very short transient events. The approximate attack times are indicated by dashed vertical lines. In such cases it is not possible to preserve the phase at the locations of multiple transients, since under modification the phase can only be preserved at one time in the life of a partial.

Strong transients are identified by the large time corrections they introduce. By breaking partials at components having large time corrections, we cause all associated par-

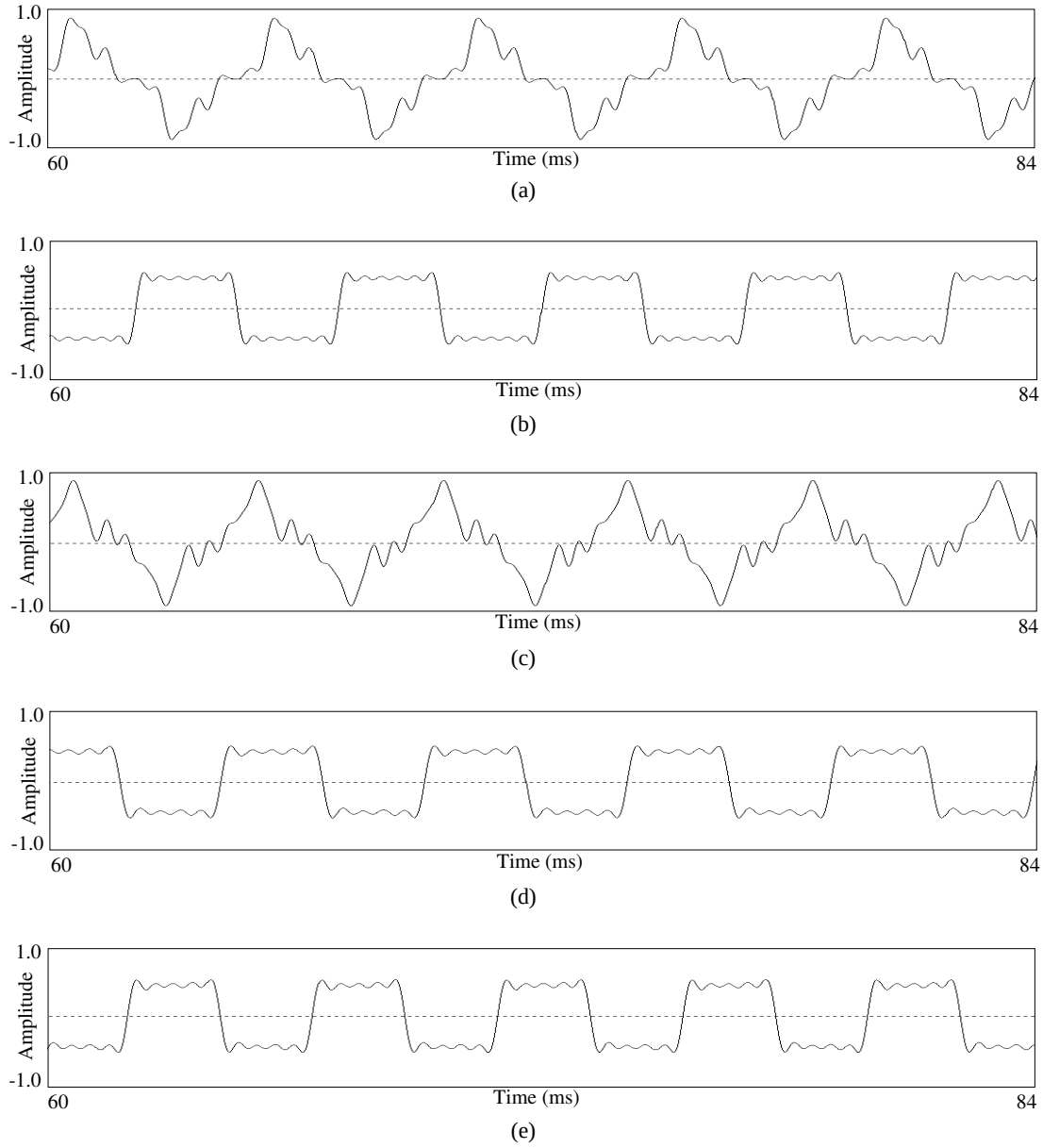


Fig. 9. Reconstruction of square wave having abrupt onset from five sinusoidal partials corresponding to first five harmonics. 24-ms plot spans slightly less than five periods of 200-Hz waveform. (a) Waveform reconstructed from nonreassigned analysis data using linear interpolation of partial frequencies. (b) Waveform reconstructed from nonreassigned analysis data using cubic phase interpolation, as proposed by McAulay and Quatieri [4]. (c) Waveform reconstructed from nonreassigned analysis data using cubic phase interpolation, with partial frequencies shifted by 10%. Notice that more periods of (distorted) waveform are spanned by 24-ms plot than by plots of unmodified reconstructions, due to frequency shift. (d) Waveform reconstructed from time–frequency reassigned analysis data using linear interpolation of partial frequencies, and having off-center components removed, or cropped. (e) Waveform reconstructed from reassigned analysis data using linear interpolation of partial frequencies and cropping of off-center components, with partial frequencies shifted by 10%. Notice that more periods of waveform are spanned by 24-ms plot than by plots of unmodified reconstructions, and that no distortion of waveform is evident.

tials to be born at the time of the transient, and thereby enhance our ability to maintain phase accuracy. In Fig. 11 the partial that spanned several transients in Fig. 10 has been broken at components having time corrections greater than the time between successive analysis window centers (about 1.3 ms in this case), allowing us to maintain the partial phases at each bongo strike. By breaking partials at the locations of transients, we can preserve the temporal envelope of multiple transient events, even under transformation.

Fig. 12(b) shows the waveform for two strikes in a bongo roll reconstructed from reassigned bandwidth-enhanced data. The same two bongo strikes reconstructed from nonreassigned data are shown in Fig. 12(a). A comparison with the source waveform shown in Fig. 12(a) reveals that the reconstruction from reassigned data is better able to preserve the temporal envelope than the reconstruction from nonreassigned data and suffers less from temporal smearing.

7 REAL-TIME SYNTHESIS

Together with Kurt Hebel of Symbolic Sound Corporation we have implemented a real-time reassigned bandwidth-

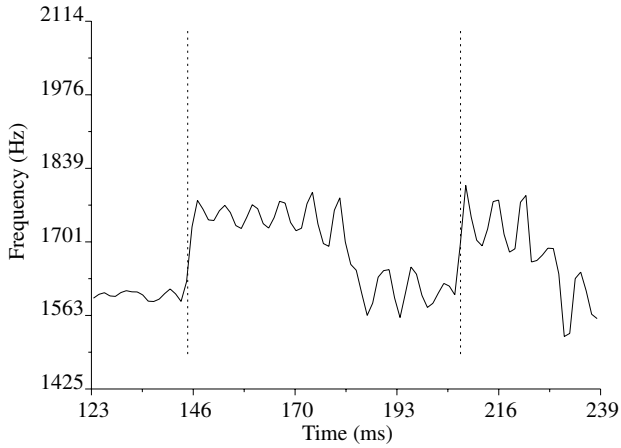


Fig. 10. Time–frequency plot of reassigned bandwidth-enhanced analysis data for one strike in a bongo roll. Dashed vertical lines show approximate locations of attack transients. Partial extends across transient boundaries. Only time–frequency coordinates of partial data are shown; partial amplitudes are not indicated.

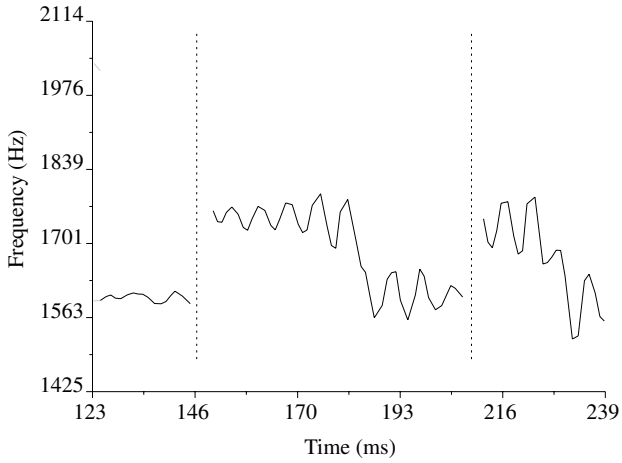


Fig. 11. Time–frequency plot of reassigned bandwidth-enhanced analysis data for one strike in a bongo roll with partials broken at components having large time corrections, and far off-center components removed. Dashed vertical lines show approximate locations of attack transients. Partials break at transient boundaries. Only time–frequency coordinates of partial data are shown; partial amplitudes are not indicated.

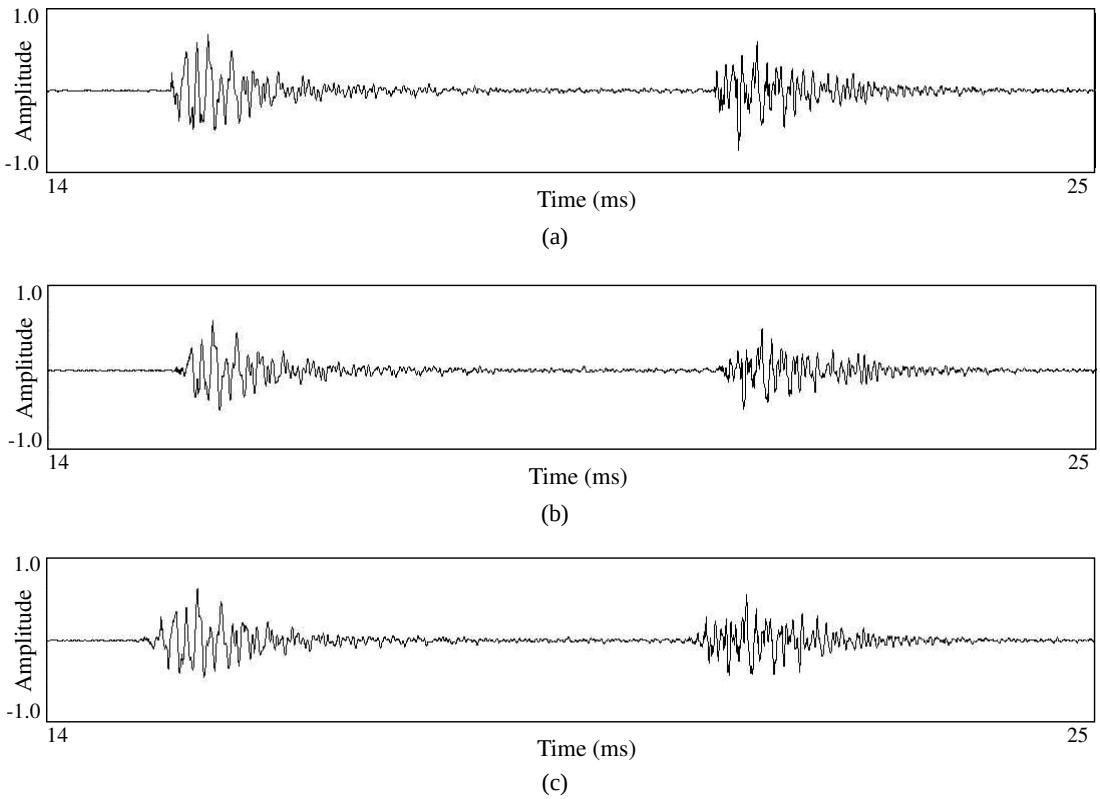


Fig. 12. Waveform plot for two strikes in a bongo roll. (a) Reconstructed from reassigned bandwidth-enhanced data. (b) Reconstructed from nonreassigned bandwidth-enhanced data. (c) Synthesized using cubic phase interpolation to maintain phase accuracy.

enhanced synthesizer using the Kyma Sound Design Workstation [15].

Many real-time synthesis systems allow the sound designer to manipulate streams of samples. In our real-time reassigned bandwidth-enhanced implementation, we work with streams of data that are not time-domain samples. Rather, our envelope parameter streams encode frequency, amplitude, and bandwidth envelope parameters for each bandwidth-enhanced partial [11], [12].

Much of the strength of systems that operate on sample streams is derived from the uniformity of the data. This homogeneity gives the sound designer great flexibility with a few general-purpose processing elements. In our encoding of envelope parameter streams, data homogeneity is also of prime importance. The envelope parameters for all the partials in a sound are encoded sequentially. Typically, the stream has a block size of 128 samples, which means the parameters for each partial are updated every 128 samples, or 2.9 ms at a 44.1-kHz sampling rate. Sample streams generally do not have block sizes associated with them, but this structure is necessary in our envelope parameter stream implementation. The envelope parameter stream encodes envelope information for a single partial at each sample time, and a block of samples provides updated envelope information for all the partials.

Envelope parameter streams are usually created by traversing a file containing frame-based data from an analysis of a source recording. Such a file can be derived from a reassigned bandwidth-enhanced analysis by resampling the envelopes at intervals of 128 samples at 44.1 kHz. The parameter streams may also be generated by real-time analysis, or by real-time algorithms, but that process is beyond the scope of this discussion. A parameter stream typically passes through several processing elements. These processing elements can combine multiple streams in a variety of ways, and can modify values within a stream. Finally a synthesis element computes an audio sample stream from the envelope parameter stream.

Our real-time synthesis element implements bandwidth-enhanced oscillators [8] with the sum

$$y(n) = \sum_{k=0}^{K-1} \left[A_k(n) + N_k(n)\, b(n) \right] \sin \theta_k(n) \qquad (12)$$

$$\theta_k(n) = \theta_k(n-1) + 2^{F_k(n)} \qquad (13)$$

where

y	=	time-domain waveform for synthesized sound
n	=	sample number
k	=	partial number in sound
K	=	total number of partials in sound (usually between 20 and 160)
A_k	=	amplitude envelope of partial k
N_k	=	noise envelope of partial k
b	=	zero-mean noise modulator with bell-shaped spectrum
F_k	=	$\log_2$ frequency envelope of partial k, radians per sample
θ_k	=	running phase for kth partial.

Values for the envelopes A_k, N_k, and F_k are updated from the parameter stream every 128 samples. The synthesis element performs sample-level linear interpolation between updates, so that A_k, N_k, and F_k are piecewise linear envelopes with segments 128 samples in length [21]. The θ_k values are initialized at partial onsets (when A_k and N_k are zero) from the phase envelope in the partial's parameter stream.

Rather than using a separate model to represent noise in our sounds, we use the envelope N_k (in addition to the traditional A_k and F_k envelopes) and retain a homogeneous data stream. Quasi-harmonic sounds, even those with noisy attacks, have one partial per harmonic in our representation. The noise envelopes allow a sound designer to manipulate noiselike components of sound in an intuitive way, using a familiar set of controls. We have implemented a wide variety of real-time manipulations on envelope parameter streams, including frequency shifting, formant shifting, time dilation, cross synthesis, and sound morphing.

Our new MIDI controller, the Continuum Fingerboard, allows continuous control over each note in a performance. It resembles a traditional keyboard in that it is approximately the same size and is played with ten fingers [12]. Like keyboards supporting MIDI's polyphonic aftertouch, it continually measures each finger's pressure. The Continuum Fingerboard also resembles a fretless string instrument in that it has no discrete pitches; any pitch may be played, and smooth glissandi are possible. It tracks, in three dimensions (left to right, front to back, and downward pressure), the position for each finger pressing on the playing surface. These continuous three-dimensional outputs are a convenient source of control parameters for real-time manipulations on envelope parameter streams.

8 CONCLUSIONS

The reassigned bandwidth-enhanced additive sound model [10] combines bandwidth-enhanced analysis and synthesis techniques [7], [8] with the time–frequency reassignment technique described in this paper.

We found that the method of reassignment strengthens our bandwidth-enhanced additive sound model dramatically. Temporal smearing is greatly reduced because the time–frequency orientation of the model data is waveform dependent, rather than analysis dependent as in traditional short-time analysis methods. Moreover, time–frequency reassignment allows us to identify unreliable data points (having bad parameter estimates) and remove them from the representation. This not only sharpens the representation and makes it more robust, but it also allows us to maintain phase accuracy at transients, even under transformation, while avoiding the problems associated with cubic phase interpolation.

9 REFERENCES

[1] F. Auger and P. Flandrin, "Improving the Readability of Time–Frequency and Time-Scale Representations by the Reassignment Method," *IEEE Trans. Signal*

Process., vol. 43, pp. 1068–1089 (1995 May).

[2] F. Plante, G. Meyer, and W. A. Ainsworth, "Improvement of Speech Spectrogram Accuracy by the Method of Spectral Reassignment," *IEEE Trans. Speech Audio Process.*, vol. 6, pp. 282–287 (1998 May).

[3] G. Peeters and X. Rode, "SINOLA: A New Analysis/Synthesis Method Using Spectrum Peak Shape Distortion, Phase and Reassigned Spectrum," in *Proc. Int. Computer Music Conf.* (1999), pp. 153–156.

[4] R. J. McAulay and T. F. Quatieri, "Speech Analysis/Synthesis Based on a Sinusoidal Representation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-34, pp. 744–754 (1986 Aug.).

[5] X. Serra and J. O. Smith, "Spectral Modeling Synthesis: A Sound Analysis/Synthesis System Based on a Deterministic Plus Stochastic Decomposition," *Computer Music J.*, vol. 14, no. 4, pp. 12–24 (1990).

[6] K. Fitz and L. Haken, "Sinusoidal Modeling and Manipulation Using Lemur," *Computer Music J.*, vol. 20, no. 4, pp. 44–59 (1996).

[7] K. Fitz, L. Haken, and P. Christensen, "A New Algorithm for Bandwidth Association in Bandwidth-Enhanced Additive Sound Modeling," in *Proc. Int. Computer Music Conf.* (2000).

[8] K. Fitz and L. Haken, "Bandwidth Enhanced Sinusoidal Modeling in Lemur," in *Proc. Int. Computer Music Conf.* (1995), pp. 154–157.

[9] T. S. Verma and T. H. Y. Meng, "An Analysis/Synthesis Tool for Transient Signals," in *Proc. 16th Int. Congr. on Acoustics/135th Mtg. of the Acoust. Soc. Am.* (1998 June), vol. 1, pp. 77–78.

[10] K. Fitz, L. Haken, and P. Christensen, "Transient Preservation under Transformation in an Additive Sound Model," in *Proc. Int. Computer Music Conf.* (2000).

[11] L. Haken, K. Fitz, and P. Christensen, "Beyond Traditional Sampling Synthesis: Real-Time Timbre Morphing Using Additive Synthesis," in *Sound of Music: Analysis, Synthesis, and Perception*, J. W. Beauchamp, Ed. (Springer, New York, to be published).

[12] L. Haken, E. Tellman, and P. Wolfe, "An Indiscrete Music Keyboard," *Computer Music J.*, vol. 22, no. 1, pp. 30–48 (1998).

[13] T. F. Quatieri, R. B. Dunn, and T. E. Hanna, "Time-Scale Modification of Complex Acoustic Signals," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing* (IEEE, 1993), pp. I-213–I-216.

[14] K. Fitz and L. Haken, "The Loris C++ Class Library," available at http://www.cerlsoundgroup.org/Loris.

[15] K. J. Hebel and C. Scaletti, "A Framework for the Design, Development, and Delivery of Real-Time Software-Based Sound Synthesis and Processing Algorithms," presented at the 97th Convention of the Audio Engineering Society, *J. Audio Eng. Soc. (Abstracts)*, vol. 42, p. 1050 (1994 Dec.), preprint 3874.

[16] M. Dolson, "The Phase Vocoder: A Tutorial," *Computer Music J.*, vol. 10, no. 4, pp. 14–27 (1986).

[17] A. Papoulis, *Systems and Transforms with Applications to Optics* (McGraw-Hill, New York, 1968), chap. 7.3, p. 234.

[18] K. Kodera, R. Gendrin, and C. de Villedary, "Analysis of Time-Varying Signals with Small *BT* Values," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-26, pp. 64–76 (1978 Feb.).

[19] D. W. Griffin and J. S. Lim, "Multiband Excitation Vocoder," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-36, p. 1223–1235 (1988 Aug.).

[20] Y. Ding and X. Qian, "Processing of Musical Tones Using a Combined Quadratic Polynomial-Phase Sinusoidal and Residual (QUASAR) Signal Model," *J. Audio Eng. Soc.*, vol. 45, pp. 571–584 (1997 July/Aug.).

[21] L. Haken, "Computational Methods for Real-Time Fourier Synthesis," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-40, pp. 2327–2329 (1992 Sept.).

[22] A. Ricci, "SoundMaker 1.0.3," MicroMat Computer Systems (1996–1997).

[23] F. Opolko and J. Wapnick, "McGill University Master Samples," McGill University, Montreal, Que., Canada (1987).

[24] E. Tellman, cello tones recorded by P. Wolfe at Pogo Studios, Champaign, IL (1997 Jan.).

APPENDIX RESULTS

The reassigned bandwidth-enhanced additive model is implemented in the open source C++ class library Loris [14], and is the basis of the sound manipulation and morphing algorithms implemented therein.

We have attempted to use a wide variety of sounds in the experiments we conducted during the development of the reassigned bandwidth-enhanced additive sound model. The results from a few of those experiments are presented in this appendix. Data and waveform plots are not intended to constitute proof of the efficacy of our algorithms, or the utility of our representation. They are intended only to illustrate the features of some of the sounds used and generated in our experiments. The results of our work can only be judged by auditory evaluation, and to that end, these sounds and many others are available for audition at the Loris web site [14].

All sounds used in these experiments were sampled at 44.1 kHz (CD quality) so time–frequency analysis data are available at frequencies as high as 22.05 kHz. However, for clarity, only a limited frequency range is plotted in most cases. The spectrogram plots all have high gain so that low-amplitude high-frequency partials are visible. Consequently strong low-frequency partials are very often clipped, and appear to have unnaturally flat amplitude envelopes.

The waveform and spectrogram plots were produced using Ricci's SoundMaker software application [22].

A.1 Flute Tone

A flute tone, played at pitch D4 (D above middle C), having a fundamental frequency of approximately 293 Hz and no vibrato, taken from the McGill University Master Samples compact discs [23, disc 2, track 1, index

3], is shown in the three-dimensional spectrogram plot in Fig. 13. This sound was modeled by reassigned bandwidth-enhanced analysis data produced busing a 53-ms Kaiser analysis window with 90-dB sidelobe rejection. The partials were constrained to be separated by at least 250 Hz, slightly greater than 85% of the harmonic partial separation.

Breath noise is a significant component of this sound. This noise is visible between the strong harmonic components in the spectrogram plot, particularly at frequencies above 3 kHz. The breath noise is faithfully represented in the reassigned bandwidth-enhanced analysis data, and reproduced in the reconstructions from those analysis data. A three-dimensional spectrogram plot of the reconstruction is shown in Fig. 14. The audible absence of the breath noise is apparent in the spectral plot for the sinusoidal reconstruction from non-bandwidth-enhanced analysis data, shown in Fig. 15.

A.2 Cello Tone

A cello tone, played at pitch D#3 (D sharp below middle C), having a fundamental frequency of approximately 156 Hz, played by Edwin Tellman and recorded by Patrick Wolfe [24] was modeled by reassigned bandwidth-

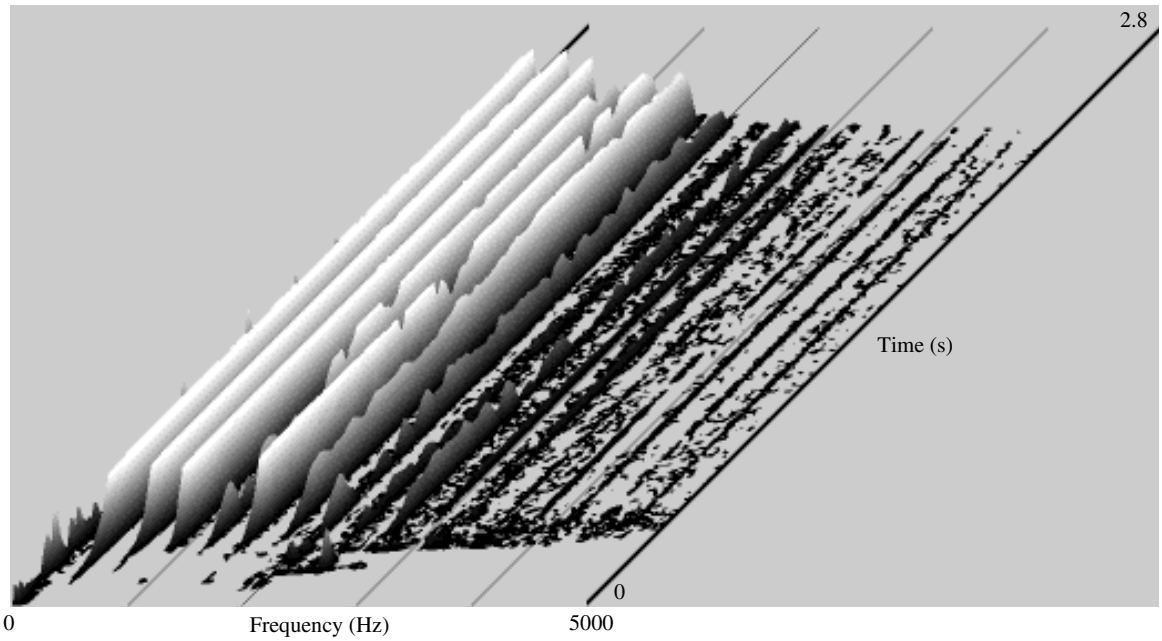


Fig. 13. Three-dimensional spectrogram plot for breathy flute tone, pitch D4 (D above middle C). Audible low-frequency noise and rumble from recording are visible. Strong low-frequency components are clipped and appear to have unnaturally flat amplitude envelopes due to high gain used to make low-amplitude high-frequency partials visible.

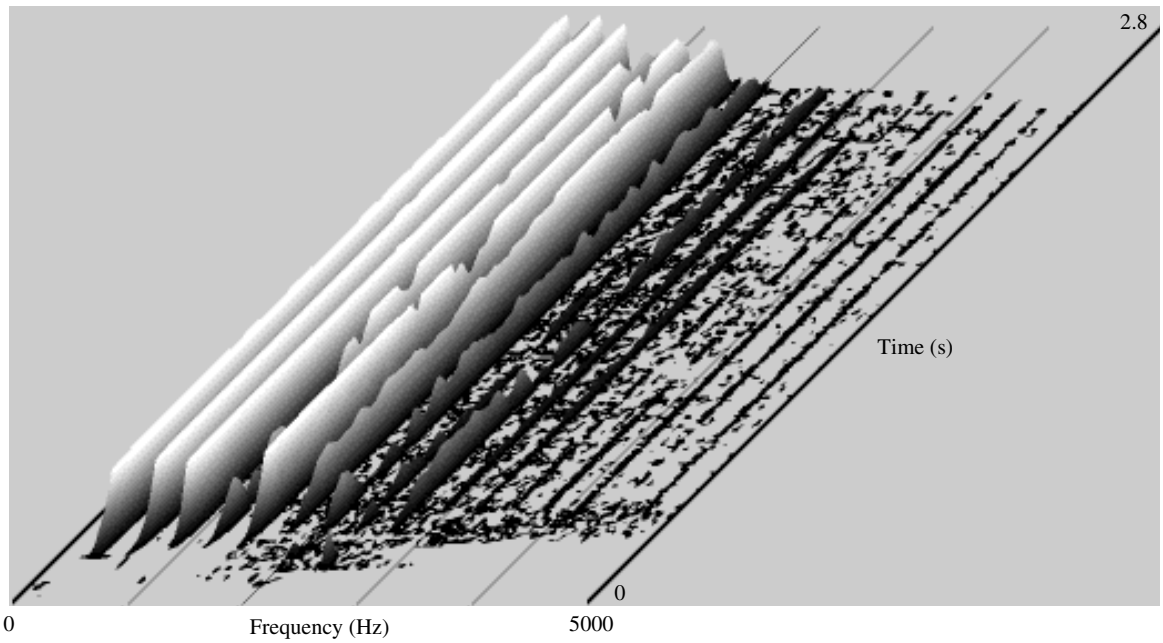


Fig. 14. Three-dimensional spectrogram plot for breathy flute tone, pitch D4 (D above middle C), reconstructed from reassigned bandwidth-enhanced analysis data.

enhanced analysis data produced using a 71-ms Kaiser analysis window with 80-dB sidelobe rejection. The partials were constrained to be separated by at least 135 Hz, slightly greater than 85% of the harmonic partial separation.

Bow noise is a strong component of the cello tone, especially in the attack portion. As with the flute tone, the noise is visible between the strong harmonic components in spectral plots, and was preserved in the reconstructions from reassigned bandwidth-enhanced analysis data and absent from sinusoidal (non-bandwidth-enhanced) reconstructions. Unlike the flute tone, the cello tone has an abrupt attack, which is smeared out in nonreassigned sinusoidal analyses (data from reassigned and nonreassigned cello analysis are plotted in Fig. 8), causing the reconstructed cello tone to have weak-sounding articulation. The characteristic "grunt" is much better preserved in reassigned model data.

A.3 Flutter-Tongued Flute Tone

A flutter-tongued flute tone, played at pitch E4 (E above middle C), having a fundamental frequency of approximately 330 Hz, taken from the McGill University Master Samples compact discs (23, disc 2, track 2, index 5], was represented by reassigned bandwidth-enhanced analysis data produced using a 17.8-ms Kaiser analysis window with 80-dB sidelobe rejection. The partials were constrained to be separated by at least 300 Hz, slightly greater than 90% of the harmonic partial separation. The flutter-tongue effect introduces a modulation with a period of approximately 35 ms, and gives the appearance of vertical stripes on the strong harmonic partials in the spectrogram shown in Fig. 16.

With careful choice of the window parameters, reconstruction from reassigned bandwidth-enhanced analysis data preserves the flutter-tongue effect, even under time dilation, and is difficult to distinguish from the original. Fig. 17 shows how a poor choice of analysis window, a 71-ms Kaiser window in this case, can degrade the representation. The reconstructed tone plotted in Fig. 17 is recognizable, but lacks the flutter effect completely, which has been smeared by the window duration. In this case multiple transient events are spanned by a single analysis window, and the temporal center of gravity for that window lies somewhere between the transient events. Time–frequency reassignment allows us to identify multiple transient events in a single sound, but not within a single short-time analysis window.

A.4 Bongo Roll

Fig. 18 shows the waveform and spectrogram for an 18-strike bongo roll taken from the McGill University Master Samples compact discs [23, disc 3, track 11, index 31]. This sound was modeled by reassigned bandwidth-enhanced analysis data produced using a 10-ms Kaiser analysis window with 90-dB sidelobe rejection. The partials were constrained to be separated by at least 300 Hz.

The sharp attacks in this sound were preserved using reassigned analysis data, but smeared in nonreassigned reconstruction, as discussed in Section 6. The waveforms for two bongo strikes are shown in reassigned and nonreassigned reconstruction in Fig. 12(b) and (c). Inspection of the waveforms reveals that the attacks in the nonreassigned reconstruction are not as sharp as in the original or the reassigned reconstruction, a clearly audible difference.

Transient smearing is particularly apparent in time-dilated synthesis, where the nonreassigned reconstruction loses the percussive character of the bongo strikes. The reassigned data provide a much more robust representation of the attack transients, retaining the percussive character of the bongo roll under a variety of transformations, including time dilation.

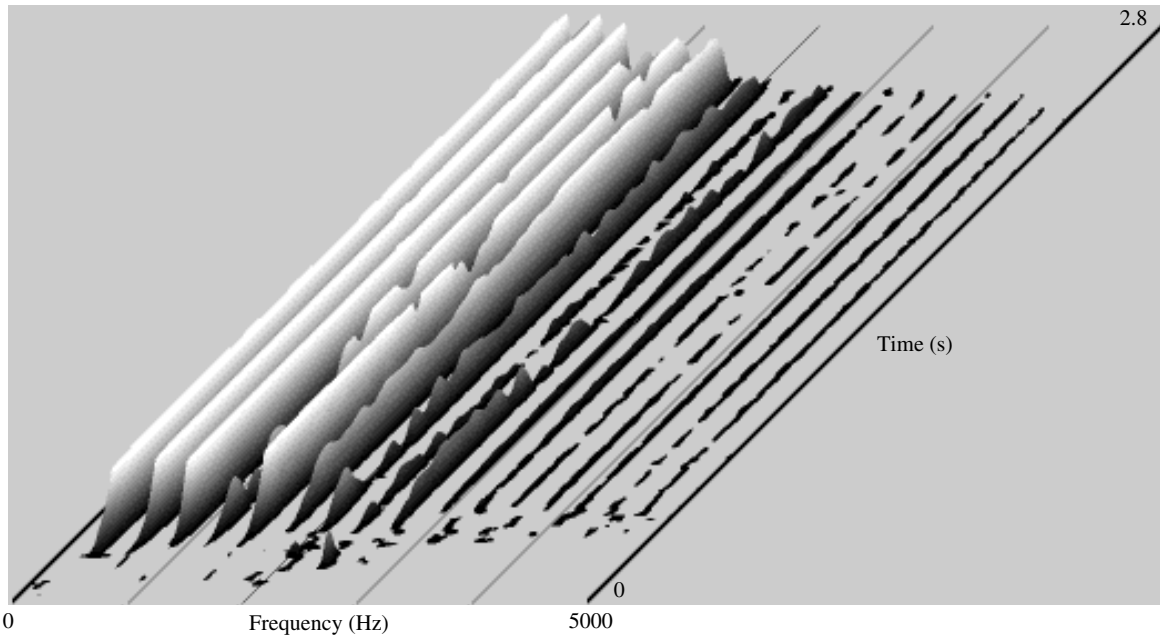


Fig. 15. Three-dimensional spectrogram plot for breathy flute tone, pitch D4 (D above middle C), reconstructed from reassigned non-bandwidth-enhanced analysis data.

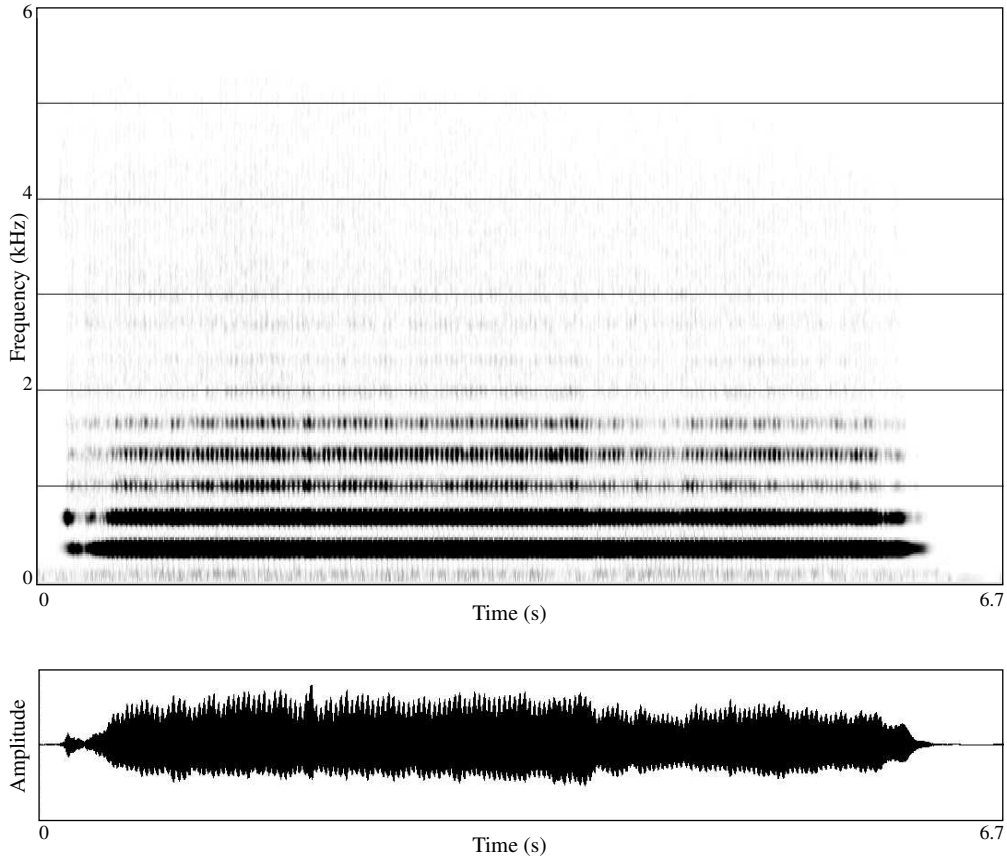


Fig. 16. Waveform and spectrogram plots for flutter-tongued flute tone, pitch E4 (E above middle C). Vertical stripes on strong harmonic partials indicate modulation due to flutter-tongue effect. Strong low-frequency components are clipped and appear to have unnaturally flat amplitude envelopes due to high gain used to make low-amplitude high-frequency partials visible.

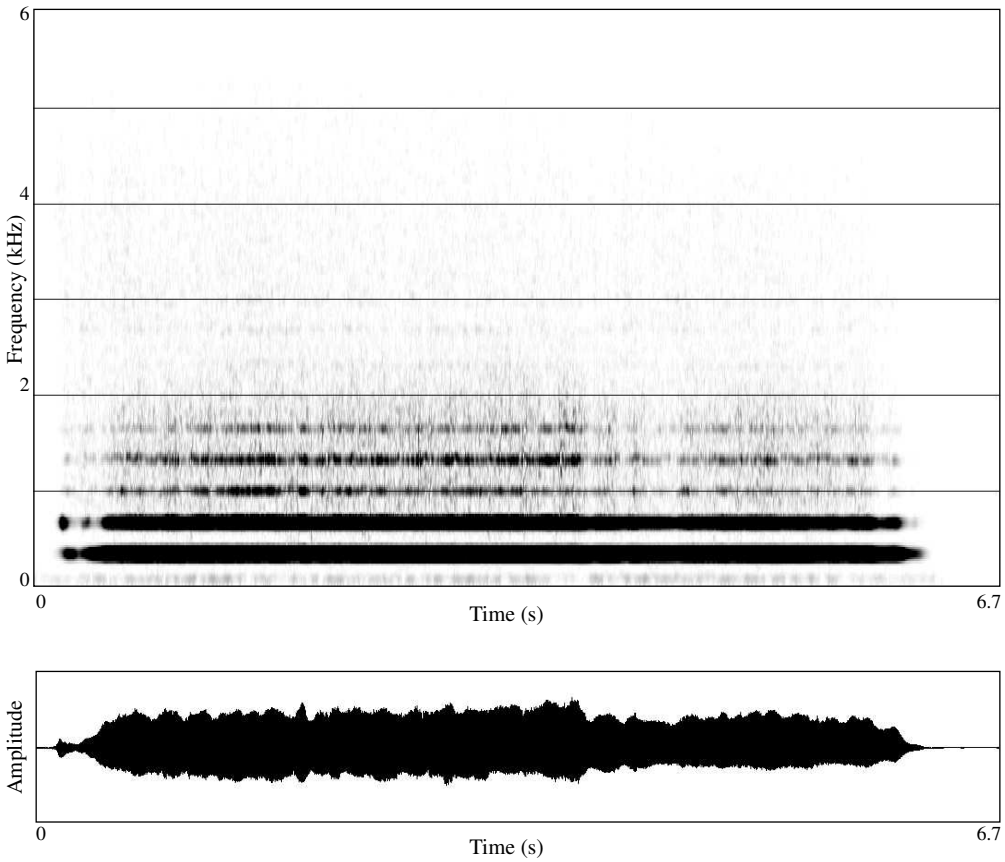


Fig. 17. Waveform and spectrogram plots for reconstruction of flutter-tongued flute tone plotted in Fig. 16, analyzed using long window, which smears out flutter effect.

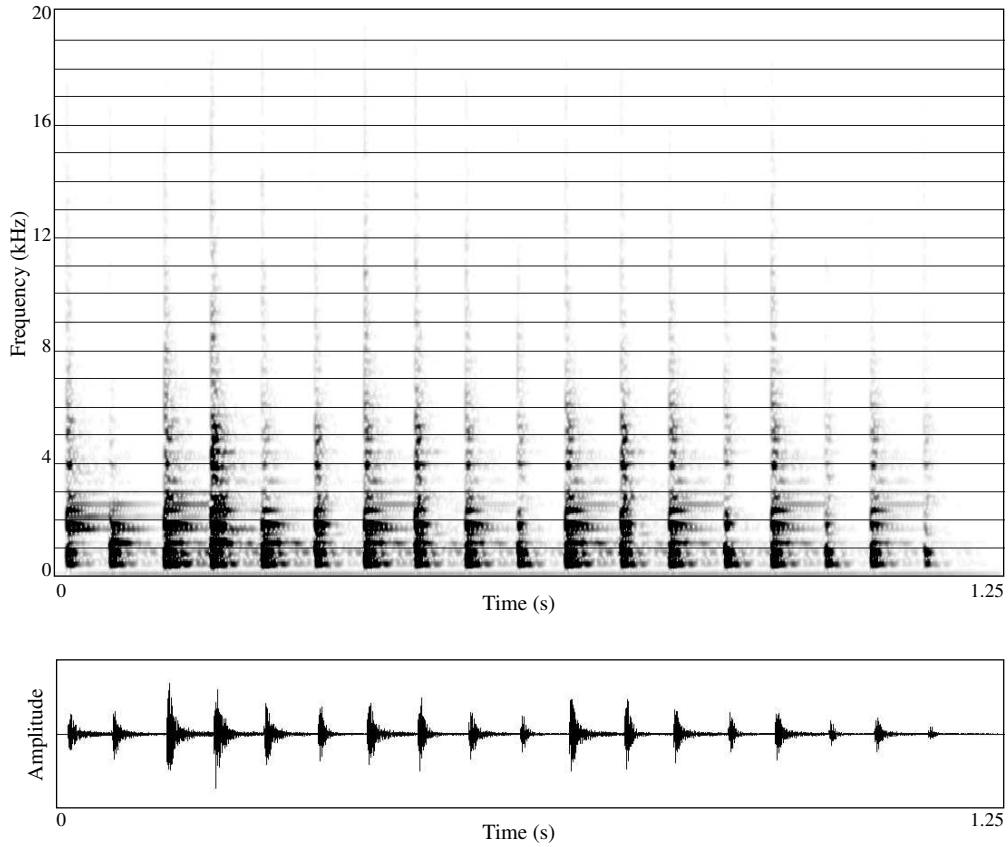
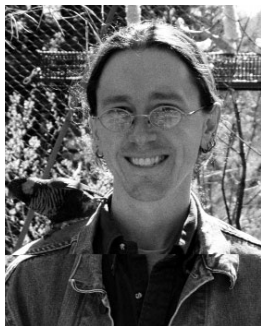


Fig. 18. Waveform and spectrogram plots for bongo roll.

THE AUTHORS



K. Fitz

L. Haken

Kelly Fitz received B.S., M.S., and Ph.D. degrees in electrical engineering from the University of Illinois at Urbana-Champaign, in 1990, 1992, and 1999, respectively. There he studied digital signal processing as well as sound analysis and synthesis with Dr. James Beauchamp and sound design and electroacoustic music composition with Scott Wyatt using a variety of analog and digital systems in the experimental music studios.

Dr. Fitz is currently an assistant professor in the department of Electrical Engineering and Computer Science at the Washington State University.

●

Lippold Haken has an adjunct professorship in electrical and computer engineering at the University of Illinois, and he is senior computer engineer at Prairie City Computing in Urbana, Illinois. He is leader of the CERL Sound Group, and together with his graduate students developed new software algorithms and signal processing hardware for computer music. He is inventor of the Continuum Fingerboard, a MIDI controller that allows continuous control over each note in a performance. He is a contributor of optimized real-time algorithms for the Symbolic Sound Corporation Kyma sound design workstation. He is also the author of a sophisticated music notation editor, Lime.

He is currently teaching a computer music survey course for seniors and graduate students in electrical and computer engineering.