MASSACHUSETTS INSTITUTE OF TECHNOLOGY
LINCOLN LABORATORY

TIME-SCALE MODIFICATION OF COMPLEX ACOUSTIC SIGNALS IN NOISE

ADA277535

T.F. QUATIERI
R.B. DUNN
R.J. McAULAY
Group 24

T.E. HANNA
Naval Submarine Medical Research Laboratory

TECHNICAL REPORT 990

4 FEBRUARY 1994

Approved for public release; distribution is unlimited.

.EXINGTON MASSACHUSETTS

This report is based on studies performed at Lincoln Laboratory, a center for research operated by Massachusetts Institute of Technology. The work was sponsored by the Naval Submarine Medical Research Laboratory under Air Force Contract F19628-90-C-0002.

This technical report has been reviewed and is approved for publication.

FOR THE COMMANDER

Gary Tutungian

Gary Tutungian
Administrative Contracting Officer
Directorate of Contracted Support Management

ABSTRACT

A new approach is introduced for time-scale modification of short-duration complex acoustic signals to improve their audibility. The method preserves an approximate time-scaled temporal envelope of a signal, thus capitalizing on the perceptual importance of the signal's temporal structure, while also maintaining the character of a noise background. The basis for the approach is a subband signal representation, derived from a filter bank analysis/synthesis, the channel phases of which are controlled to shape the temporal envelope of the time-scaled signal. Channel amplitudes and filter bank inputs are selected to shape the spectrum and correlation of the time-scaled background. The phase, amplitude, and input control are derived from locations of events that occur within filter bank outputs. A frame-based generalization of the method imposes phase consistency and background noise continuity across consecutive synthesis frames. The approach and its derivatives are applied to synthetic and actual complex acoustic signals consisting of closely spaced sequential time components.

ACKNOWLEDGMENTS

We acknowledge John Harvey of the Naval Submarine Medical Research Center for many hours of constructive feedback, which has helped shape our directions and improve our enhancement algorithms. We also thank the Lincoln Laboratory Innovative Research Program Technical Review Committee for supporting Tom Quatieri and Bob Dunn for a major component of this work.

TABLE OF CONTENTS

TABLE OF CONTENTS
(Continued)

LIST OF ILLUSTRATIONS

1. INTRODUCTION

Short-duration complex sounds, as from the closing of a stapler or the tapping of a drum stick, often consist of a series of brief components that are closely spaced and possibly overlapping in time. Consequently, the sound components are difficult to detect aurally, and two such similar acoustic sounds are difficult to discriminate. The method of enhancing complex acoustic signals introduced in this report is time-scale expansion, sometimes referred to as "slow-motion audio replay," which extends the time scale of the signal without altering its spectral character. This expansion allows the listener to capture tne sequential components of short-duration complex signals and so has a wide variety of applications, including underwater, biological, and music sound processing. The complex signal is also typically embedded in noise. The signal to be processed is given by $s(t) = x(t) + b(t)$, where $x(t)$ is the signal of interest, or object, and $b(t)$ is the noise, or background. The object may consist of sequential events such as rapidly damped sine waves, short-duration steady tones, or noise bursts. The background is generally nonstationary and may consist of deterministic as well as random components.

Any approach to time-scale modification of such signals must satisfy two constraints for high-fidelity reconstruction. The first stems from the observation that complex acoustic sounds are characterized by a rich temporal envelope structure that appears to play an important role in auditory discrimination. In many nonspeech sounds [1–3], as well as unvoiced speech sounds [4], the temporal envelope has been found to be a perceptual cue for recognition. Enhancement of these sounds should account for the perceptual importance of the temporal envelope. The second constraint is preservation of the background, i.e., the time-scaled background should have the same character as the original so that it will not be misconstrued as an object component. The goal here is to time-scale modify $s(t)$ to improve the audibility of the object $x(t)$ without changing the character of the background $b(t)$.

Methods of time-scale modification can smear the fine structure of the temporal envelope. This result is true even for high quality techniques developed in the speech context such as the phase vocoder [5] and sine-wave analysis/synthesis [6,7]. Loss of temporal resolution results from windowing the signal during analysis and dispersing the phase during synthesis. Windowing dulls the sound and merges components while phase dispersion tends to introduce a reverberant, tinny quality. In addition, current methods of time-scale modification can introduce artifacts into the background, e.g., unwanted tonality or "gurgles." This effect can be due to excessive correlation of sine-wave components in time [6,7] or spectral corruption in frequency [5]. This report introduces a new approach for time-scale modification of complex acoustic signals where the object consists of sums of sequential, rapidly damped sine waves. While maintaining the character of the noise background, the method preserves an approximate time-scaled temporal envelope of the signal, thus capitalizing on the perceptual importance of a signal's temporal structure. Derivatives of the approach are used to address a more general object class.

The framework for the new approach is a subband signal representation that is derived from a filter bank analysis/synthesis. The essence of the approach is to shape the temporal envelope of the time-scaled object component by controlling relative phases of subband components and to provide the appropriate spectral shaping and correlation to the time-scaled background by controlling the subband amplitudes and filter bank inputs. The subband phase and amplitude and filter bank input control are derived from locations of object events in time and frequency. A frame-based generalization of the method imposes phase consistency and background noise continuity across consecutive synthesis frames. Improved audibility of the object was achieved without annoying background artifacts for a large range of acoustic signals.

The outline of the report is as follows: Section 2 presents the subband signal representation; Section 3 uses the framework of Section 2 as the basis to time-scale modify signals consisting of sums of sequential, rapidly damped sine waves. Section 4 applies the subband framework to time-scale modification of background; the subband approaches to modification of object and background are then combined into a single filter bank-based system. Section 5 describes a strawman design for a more general object class. This system is applied to underwater sounds in a variety of signal-to-noise conditions. Section 6 summarizes and describes current efforts.

2. A SUBBAND FRAMEWORK

Motivation for the subband framework is threefold. First, filter bank analysis/synthesis can be formed as an identity system, thus giving the potential for representing fine temporal and spectral structure. Second, the time-frequency distribution corresponding to the subband representation provides a means for determining the presence or absence of an object component. And third, the subband decomposition is compatible with a sine-wave representation of signals [8]; hence the features of the two approaches can be applied synergistically.

2.1 Subband Representation

The subband signal representation is formulated as follows. Consider a discrete-time signal $x(n)$[1] passed through a bank of filters $h_k(n)$ where each filter is given by a modulated version of a baseband prototype filter $h(n)$, i.e.,

$$h_k(n) = h(n)\exp[j(2\pi/R)kn] \quad , \tag{1}$$

where $h(n)$ is assumed causal and lies over a duration $0 \leq n < S$, and R is the frequency sampling factor. The filters are designed to satisfy a perfect reconstruction constraint in frequency

$$\sum_k h_k(n) = \delta(n) \quad , \tag{2}$$

where $\delta(n)$ is the unit sample sequence. One condition for perfect reconstruction is that the length of $h(n)$ be less than the frequency sampling factor, i.e., $S < R$ [9]. Each channel output $y_k(n)$ is given by

$$y_k(n) = x(n) * h_k(n) \tag{3}$$

and with the perfect reconstruction constraint ($S < R$), the signal $x(n)$ can be recovered as

$$\begin{aligned} y(n) &= \sum_k x(n) * h_k(n) \\ &= x(n) * \sum_k h_k(n) \end{aligned}$$

[1] Because the signal processing is performed digitally, sampled data notation will be used throughout this report.

$$= x(n) \quad . \tag{4}$$

Each filter output $y_k(n)$ is complex [each filter response $h_k(n)$ in Equation (1) is complex] so that the temporal envelope of the output of the kth channel is

$$a_k(n) = |y_k(n)| \quad , \tag{5}$$

and the phase of each bandpass output is

$$\theta_k(n) = \tan^{-1}(Im[y_k(n)]/Re[y_k(n)]) \quad . \tag{6}$$

Thus the output of each filter can be viewed as an amplitude and phase modulated (complex) sine wave

$$y_k(n) = a_k(n)\exp[j\theta_k(n)] \quad , \tag{7}$$

and reconstruction of the signal can be viewed as a sum of complex exponentials

$$x(n) = \sum_k a_k(n)\exp[j\theta_k(n)] \quad , \tag{8a}$$

where the amplitude and phase components are given in Equations (5) and (6). Because the filters are hermetian symmetric with respect to the filter index k [8], the complex expression in Equation (8a) can be reduced to a sum of real sine waves of the form

$$x(n) = \sum_k b_k(n)\cos[\theta_k(n)] \quad , \tag{8b}$$

where $b_k(n)$ is a scaled representation of $a_k(n)$.

2.2 Short-Time Representation

Because the signal processing occurs on a short-time basis, the signal $x(n)$ is windowed with a sequence $w(n)$ of length N. This window is repeatedly applied at some frame interval L to create the short-time segments

$$z_l(n) = w(n - lL)x(n) \quad , \tag{9a}$$

where the sliding window is designed for perfect reconstruction in time

$$\sum_l w(n - lL) = 1 \quad . \tag{9b}$$

Each short-time segment $z_l(n)$, when passed through the filter bank $h_k(n)$, can thus be represented by a sum of complex sine waves of the form in Equation (7):

$$\begin{aligned} y_{k,l}(n) &= h_k(n) * z_l(n) \\ &= a_k^l(n)\exp[j\theta_k^l(n)] \quad . \end{aligned} \tag{10}$$

Given the time-frequency identities from Equations (2) and (9b), when the filter outputs of all segments $y_{k,l}(n)$ are summed, the result is the original sequence $x(n)$:

$$\begin{aligned} y(n) &= \sum_l \sum_k y_{k,l}(n) \\ &= \sum_l \sum_k h_k(n) * z_l(n) \\ &= \sum_l \sum_k h_k(n) * w(n - lL)x(n) \\ &= x(n) \sum_l \sum_k h_k(n) * w(n - lL) \\ &= x(n) \sum_l w(n - lL) * \sum_k h_k(n) \\ &= x(n) \sum_l w(n - lL) * \delta(n) \\ &= x(n) \sum_l w(n - lL) \\ &= x(n) \quad . \end{aligned} \tag{11}$$

This identity, illustrated in Figure 1, represents the subband (and overlap-add) framework for the approaches developed within this report.

2.3 The Problem of Time-Scale Modification

One approach to time-scale modification relies on the subband signal representation in Equation (8a). The output of each filter is viewed as an amplitude- and phase-modulated sine wave,

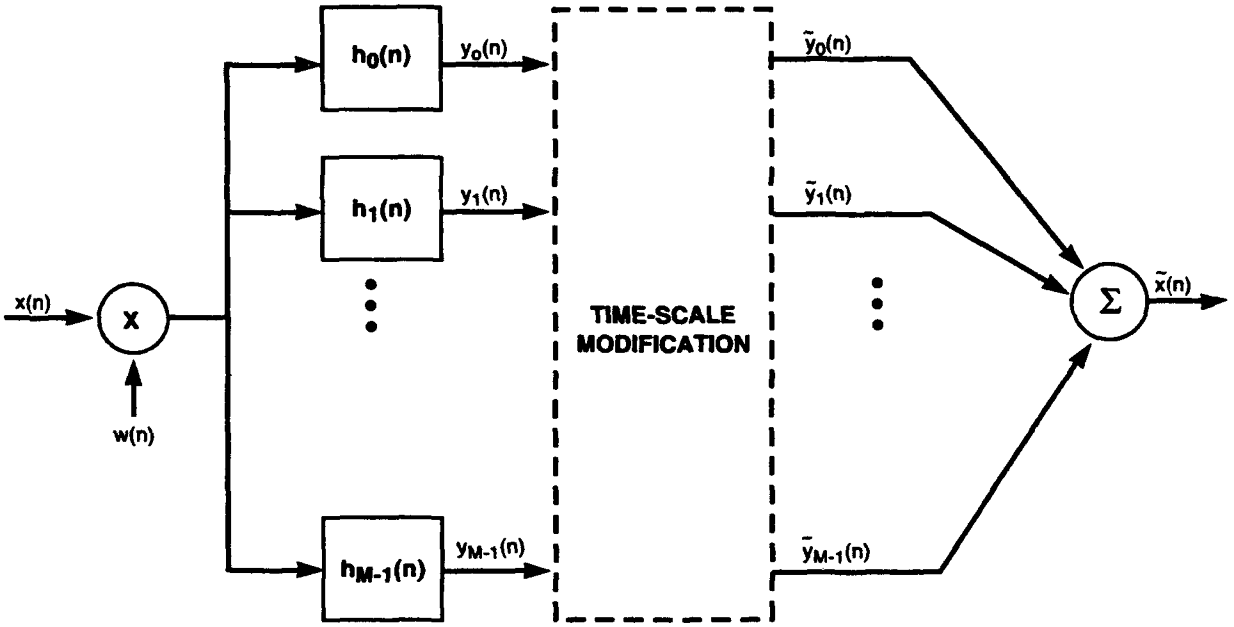

Figure 1. Filter bank/overlap-add reconstruction with time-scale modification.

the amplitude and unwrapped phases of which are interpolated to perform time-scale modification (see Figure 1). With time-scale modification by a factor ρ, the modified filter output is given by[2]

$$\tilde{y}_k(n) = \tilde{a}_k(n)\cos[\rho\tilde{\theta}_k(n)] \quad , \tag{12}$$

where $\tilde{a}_k(n)$ is the channel envelope and $\tilde{\theta}_k(n)$ is the unwrapped phase, both interpolated by the factor ρ. The interpolated phase function in Equation (12) is scaled by ρ to maintain the original frequency trajectory (i.e., phase derivative) of each filter output. This technique, which is the approach to time-scale modification used in the phase vocoder, generally distorts the temporal envelope of short-duration complex signals [5]. Because the original phase relation among channels is lost through the phase modification in Equation (12), the temporal envelope of the composite signal $\sum_k \tilde{y}_k(n)$ generally will not be a time-scaled version of the original signal envelope. Although with interpolation the shape of the temporal envelope of each filter output can be preserved, the envelope of the composite signal may differ significantly from the original. (An example of this

[2] For convenience the real representation of the complex exponentials are used.

distortion is shown in Section 3.) This approach has the additional problem that background noise is distorted in the form of gurgles or tonality, the nature of the distortion being a function of the prototype filter $h(n)$ and the window $w(n)$.

2.4 Relation with a Sine-Wave Representation

The subband representation of Equation (8a) bears a strong resemblance to a sine-wave representation originally developed in the speech context [7]. In this representation the signal is modeled as a sum of sine waves with time-varying amplitudes, frequencies, and phases:

$$y(n) = \sum_{k} A_k(n) \cos[\Theta_k(n)] \quad . \tag{13}$$

The time-varying frequency of each sine wave is given (in continuous time) by the derivative of the phase and denoted by $\omega_k(t) = \dot{\Theta}_k(t)$. The sine-wave parameters are estimated by spectral peak-picking using short-time Fourier analysis. The synthesis is performed using amplitude and phase interpolation and sine-wave summation. Although this model was originally formulated for speech signals, it is also capable of representing nonspeech signals such as underwater, biological, and music sounds.[3]

As in the phase vocoder, a time-scaled modified signal can be formed using a modification similar to Equation (12). The advantage of this system is that it is well suited for time-scale modification of tone-like signals, because it explicitly models and estimates sine waves. Appendix A reviews and further develops a sine-wave modification system and shows, along with discussions throughout this report, that the synergism between the sine-wave and subband approaches to time-scale modification is coming full circle. Many of the features developed in the sine-wave framework to refine time-scale modification and coding applications in the speech context, such as phase synchronization in voiced speech [9] and phase dithering in unvoiced speech [10], have been stepping stones to the subband approach developed in this report. Likewise the methods newly developed in the subband framework are helping to improve sine-wave-based modification.

[3] Although the sine-wave representation can provide a near perceptual identity for most nonspeech sounds, it is not a strict identity.

3. TEMPORAL ENVELOPE INVARIANCE

The temporal envelope of a signal reflects its distinctive time-domain features, e.g., the start and stop time of a click, the sharp attack or slow decay of a damped sine wave, or the modulation pattern of two beating sine waves. Temporal envelope is sometimes defined, typically in the context of bandpass signals, as the magnitude of the corresponding analytic signal representation derived from the Hilbert transform [11]. Other definitions of temporal envelope have been proposed based on estimates of attack and release dynamics [12]. One approach to time-scale modification (developed in Appendix B), given the spectral envelope of a signal, is to select a Fourier-transform phase that results in a sequence with a time-scaled version of the original temporal envelope. A close match to both the spectral envelope and modified temporal envelope, however, may not be consistent with the relationship between a sequence and its Fourier transform. For example, expanding the temporal envelope of an exponentially damped sine wave by slowing the decay rate narrows the signal's resonant bandwidth. Arbitrarily slowing the decay rate and maintaining the original resonant bandwidth violates constraints on the signal's time and frequency concentrations [13]. In general, there may not necessarily exist a sequence jointly satisfying the desired temporal and spectral envelope constraints. Therefore, the signal modification problem is formulated (in Appendix B) as finding a fullband Fourier-transform phase that results in a sequence with a temporal envelope that is close in some sense to a desired time-scaled envelope.[4]

This section develops an approach to preserving an approximate temporal envelope in time-scale modification in the context of a subband signal representation. The objective of the approach then is similar to the fullband technique; however, the essence of the technique is to maintain the phase relation of channels at time instants that are associated with distinctive features of the envelope [14,15] rather than explicitly attempting to maintain the temporal envelope over all time. As a stepping stone to the approach, the notion of *instantaneous invariance* is introduced. Subband temporal envelopes are then used to control the shape of the fullband temporal envelope.

3.1 Instantaneous Invariance

It is assumed that the temporal envelope of a waveform near a particular time instant $n = n_o$ is determined by the amplitude and phase of its subband components at that time [i.e., $a_k(n_o)$ and $\theta_k(n_o)$], and by the time rate of change of these amplitude and phase functions. To preserve the temporal envelope in the new time scale near $n = \rho n_o$, these amplitude and phase relations are

[4] The fullband approach of Appendix B attempts to preserve the time-scaled version of the signal temporal envelope by iteratively selecting a best set of Fourier transform phases. Potential problems that arise with this approach are the computational complexity (sometimes requiring hundreds of iterations) and, more importantly, the difficulty in obtaining a meaningful fullband envelope through the analytic signal representation.

maintained at that time (Figure 2). Interpolation of amplitude and frequency as in Equation (12) does not maintain the phase relations; however, it does maintain the amplitudes and amplitude and phase derivatives. The phase relations can be maintained by adding (to each channel phase) an offset, guaranteeing that the corresponding integrated frequency trajectory takes on the desired phase at the specified time $n = \rho n_o$. Introduced in each channel is a phase correction that sets the phase of the modified filter output $\tilde{y}_k(n)$ at $n = \rho n_o$ to the phase at $n = n_o$ in the original time scale. Denoting the phase correction by ϕ_k, the modified channel signal becomes

$$\tilde{y}_k(n) = \tilde{a}_k(n)\cos[\rho\tilde{\theta}_k(n) + \phi_k] \quad , \tag{14a}$$

where

$$\phi_k = \theta_k(n_o) - \rho\tilde{\theta}_k(\rho n_o) \quad . \tag{14b}$$

This method of phase synchronization was inspired by an approach to time-scale modification of voiced speech that uses glottal onset times [9].

An inconsistency arises, however, when preservation of the temporal envelope is desired at more than one time instant. Figure 3 illustrates the eloquence and the difficulty of the approach by showing that when the phase offsets are selected to preserve the temporal envelope at one time instant, there is no guarantee that the envelope will be maintained elsewhere. (Filter bank specifications are given in Section 3.2.) One approach to resolving this inconsistency is to allow specific groups of subband components to contribute to different instants of time at which invariance is desired.

3.2 Invariance in Clustered Subbands

The approach to invariance can be described by using the signal (in Figure 3) that has a high- and low-frequency component, each with a different onset time. If all the channels are phase-aligned near the low-frequency event, the phase relations at the high-frequency event are changed and vice versa. For this signal with two events of different frequency content, it is preferable to distribute the phase alignment over the two events—the high-frequency channels being phase-aligned at the first event and the low-frequency channels being phase-aligned at the second event. Equation (14a) can then be applied to each channel group using the time instant for the respective event, thus aligning the channel phases that most contribute to each event.

Channels are assigned to time instants using the temporal envelope of the filter bank outputs. Accordingly, the filter bank is designed such that each filter output reflects distinctive events that characterize the temporal envelope of the input signal. This constraint requires filters of short duration. The filters must also be smooth in time so that the temporal envelope of each filter output will not exhibit spurious peaks due to the filter itself. To meet these requirements a perfect

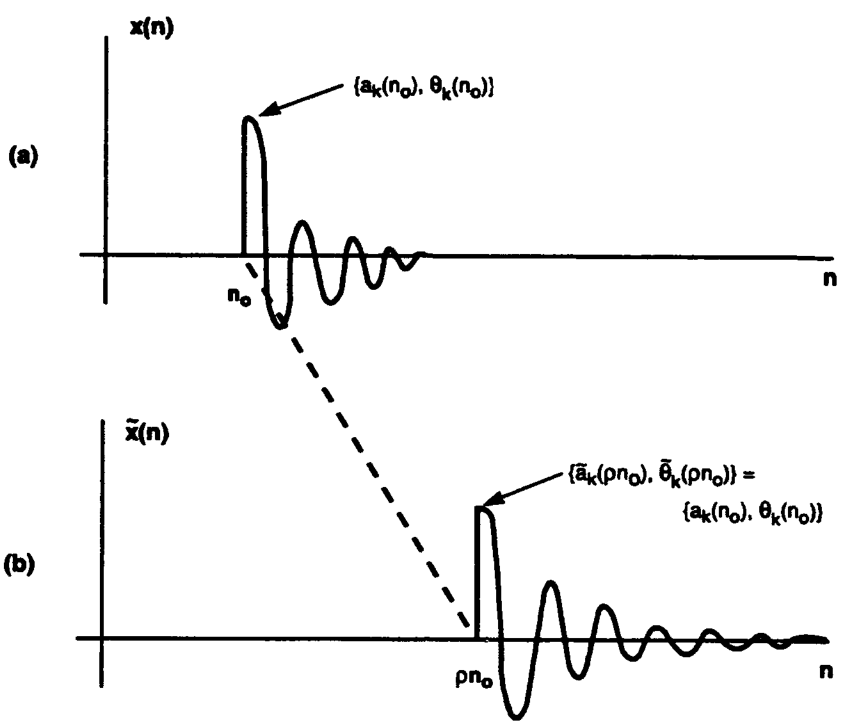
x(n)
$\{a_k(n_o), \theta_k(n_o)\}$
(a)
n_o
n
$\tilde{x}(n)$
$\{\tilde{a}_k(\rho n_o), \tilde{\theta}_k(\rho n_o)\} =$
$\{a_k(n_o), \theta_k(n_o)\}$
(b)
ρn_o
n

Figure 2. Invariance of subband amplitudes and phases at time instant $n = n_o$: (a) original and (b) time-scaled.

reconstruction filter bank with 21 uniformly spaced filters $h_k(n)$ was designed using a prototype Gabor filter (i.e., of Gaussian shape). This prototype had an effective duration of about 2 ms and was truncated to meet the perfect reconstruction constraint that its length be less than the frequency sampling factor. Using the channel envelopes derived from the filter bank, channels are clustered according to their similarity in the temporal envelope across frequency. This approach was inspired by the work of Mallat in a wavelet representation of waveform singularities [16].

The clustering is performed as follows. First, the *occurrence time of an event* is defined within each channel as the location of the maximum of $a_k(n)$ and denoted by $n_o(k)$. It is assumed that the signal is of short duration with no more than two events and that only one occurrence time is assigned to each channel; more generally, multiple occurrence times may be required. An interesting property of the occurrence times $n_o(k)$ is that they tend to cluster near singularities such as sharp attacks or decays. This observation is similar to that made by Mallat for nonuniform filter banks [16]. A histogram analysis of $n_o(k)$ shows the clustering and forms the basis for event selection. In particular, a histogram of occurrence times is formed, and the average values of each of the two

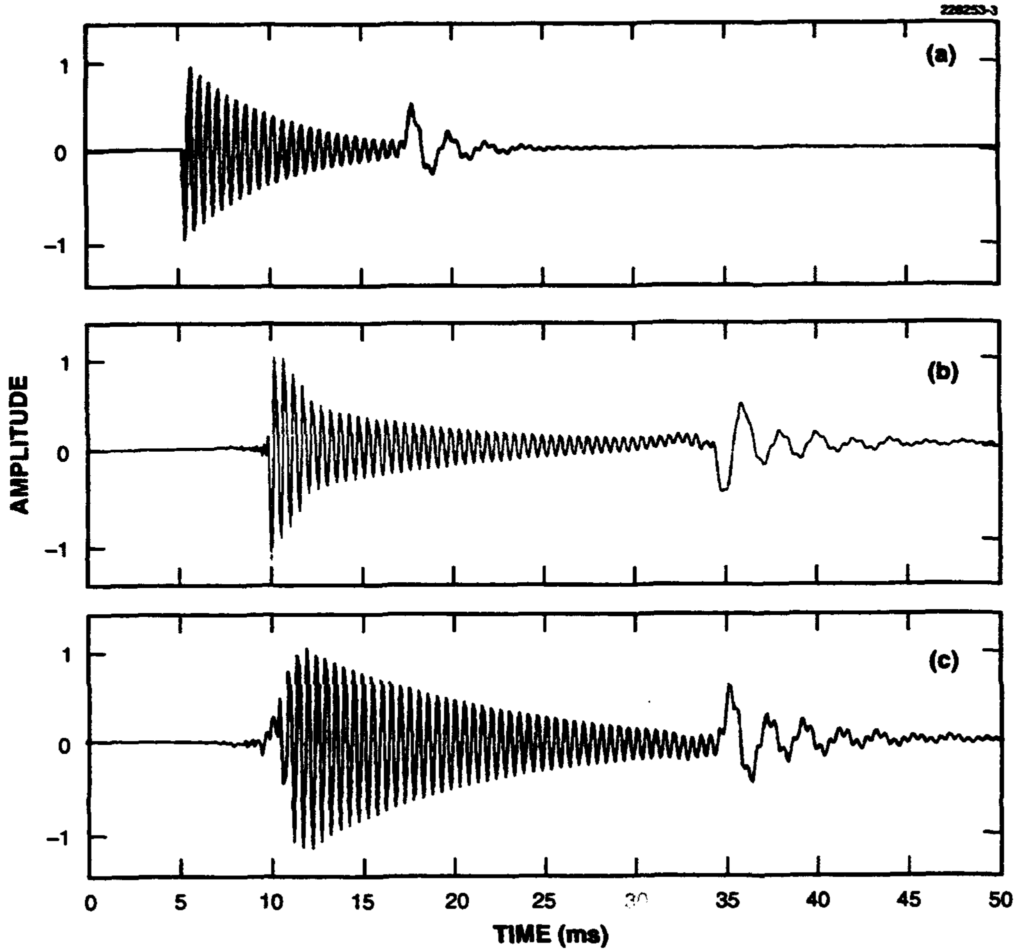

Figure 3. Time-scaled expansion (×2) using subband phase correction: (a) original and (b) expansion with left and (c) with right invariance.

highest bins are selected as the event locations. These times are denoted by n_o^1 and n_o^2, and each of the k channels is assigned to n_o^1 or n_o^2 based on the minimum distance between $n_o(k)$ and the two possible event time instants. The distance is given by

$$D(p) = |n_o(k) - n_o^p| \quad , \tag{15}$$

where $p = 1, 2$. The resulting two clusters of channels are denoted by $\{y_k^p(n)\}$ with $p = 1, 2$.

Finally, based on the channel assignment a phase correction is introduced in each channel, making the phase of the modified filter output $\tilde{y}_k^p(n)$ at time $n = \rho n_o^p$ equal to the phase at the time instant $n = n_o^p$ in the original time scale. Denoting the phase correction for each cluster by ϕ_k^p, the modified channel signal becomes

$$\tilde{y}_k^p(n) = \tilde{a}_k^p(n)\cos[\rho\tilde{\theta}_k^p(n) + \phi_k^p] \quad , \tag{16a}$$

where

$$\phi_k^p = \theta_k^p(n_o^p) - \rho\tilde{\theta}_k^p(\rho n_o^p) \quad , \tag{16b}$$

and where p refers to the first or second cluster. The original phase relations within each channel cluster will then be preserved at the assigned time instant.

The filter bank approach is illustrated in Figure 4, where the signal in Figure 3 has been passed through the Gabor filter bank with the phase correction of Equation (16b). Figure 4(a) shows the maxima of the temporal envelope of the filter outputs as a two-dimensional function of time and filter number, and it demonstrates that the maxima cluster near the onset times of the two damped sine waves. The time-scaled (×2) signal shown in Figure 4(b) approximately preserves a time-scaled version of the original temporal envelope. An expansion of the signal without phase correction, illustrated in Figure 4(c), significantly modifies the temporal envelope. The filter bank approach is also demonstrated in Figure 5, which shows modification of an actual acoustic signal from a closing stapler.

In these examples, as well as with a variety of other signals, the components of the acoustic signal that were barely audible in the original are enhanced through time expansion. Also observed was a slight narrowing of resonant bandwidths that manifests itself perceptually as a mild sharpening of the sound. Given the fidelity of the expanded temporal envelope and the constraints on a signal's time and frequency concentrations [13], this effect is expected.

3.3 Short-Time Processing of Long Waveforms

To process a long waveform, the signal is segmented as in Equation (9a), and the filter bank-based method is applied to each segment. Synthesis is performed by overlapping and adding

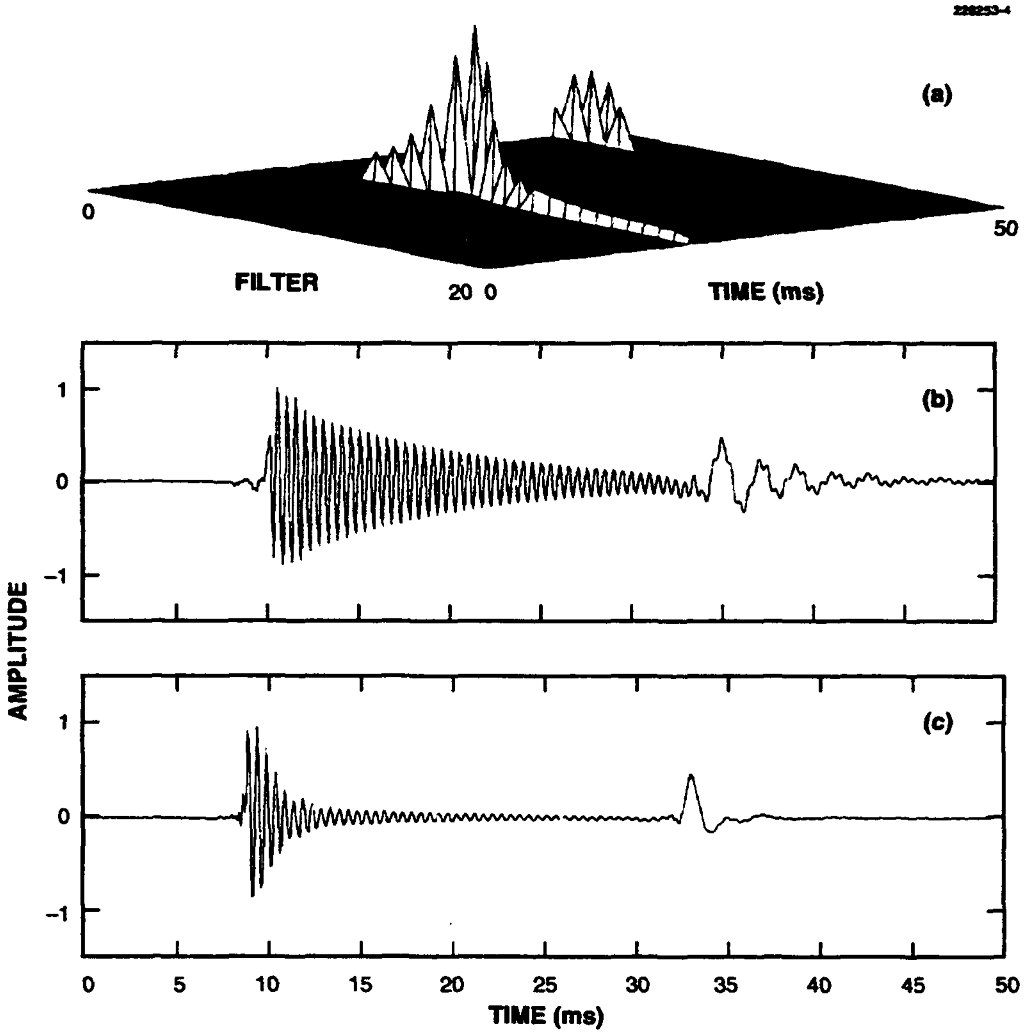

Figure 4. Time-scaled expansion (×2) using clustered subband phase correction: (a) envelope maxima and (b) expansion with and (c) without phase control.

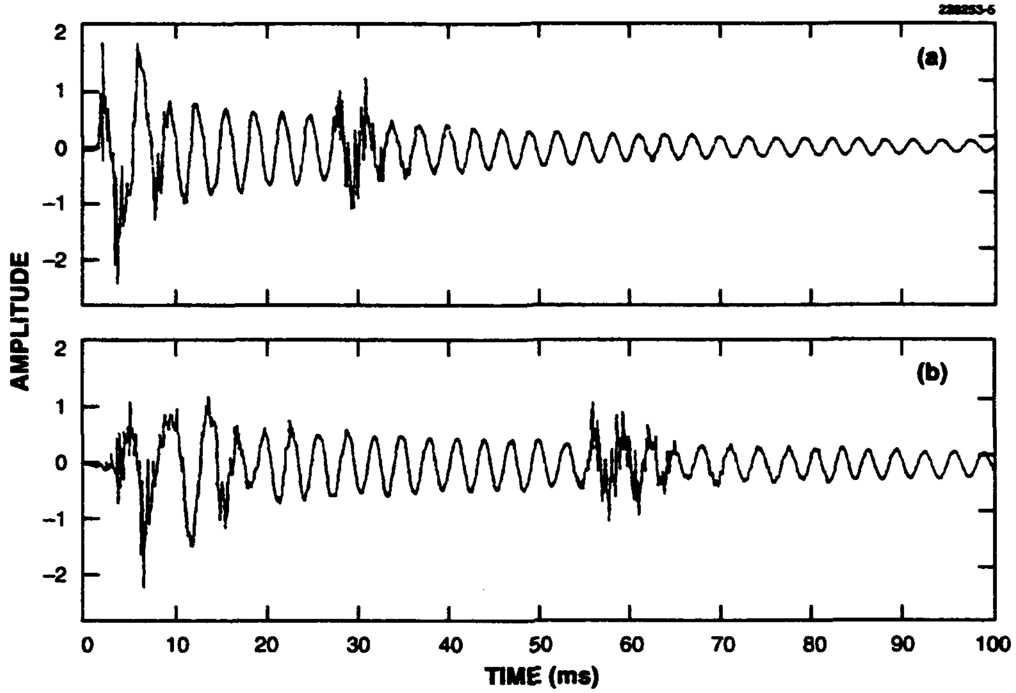

Figure 5. Time-scaled expansion (×2) of a response from a closing stapler: (a) original and (b) time-scaled.

adjacent time-scaled segments, and consistency is accounted for between any events common to adjacent frames.[5] If no event is detected, then a smooth phase transition is made across frame boundaries.

Specifically, the lth segment is denoted by $z_l(n) = w(n-lL)x(n)$, where $w(n)$ is a sliding analysis window and L, the frame interval, is set to half the window length. The window $w(n)$ is chosen such that $\sum_l w(n-lL) = 1$, i.e., the overlapping windows form an identity as in Equation (9b). The first step in the algorithm is to apply the filter bank modification to the windowed segment $z_l(n)$. The time instants of the events are saved and time-normalized with respect to the next

[5] The notion of synchronizing pulses over adjacent frames in overlap-add for time-scale modification was first introduced by Roucos and Wilgus in the speech context [17]. This method relies on cross-correlation of adjacent frames to align pulses and not on a subband decomposition or phase synchronization.

frame. The procedure is repeated for frame $l+1$. However, if the most recent event from frame l falls at least $L/4$ samples inside the current frame $l+1$, then this event is designated the first event of frame $l+1$. Under this condition, the second event is found via the maximum of the histogram of the event occurrence times on frame $l+1$. This case is illustrated in Figure 6. As in the above procedure, each channel is then assigned to a time instant based on the two event times and the measured occurrence times $n_o(k)$. In addition, a frame is also allowed to have no events by setting a histogram bin threshold below which a no-event condition is declared (typically, a threshold of two or three within a cluster). In this case, channel phase offsets are selected to make the channel phases continuous across frame boundaries, i.e., the phase is allowed to "coast" from the previous frame.

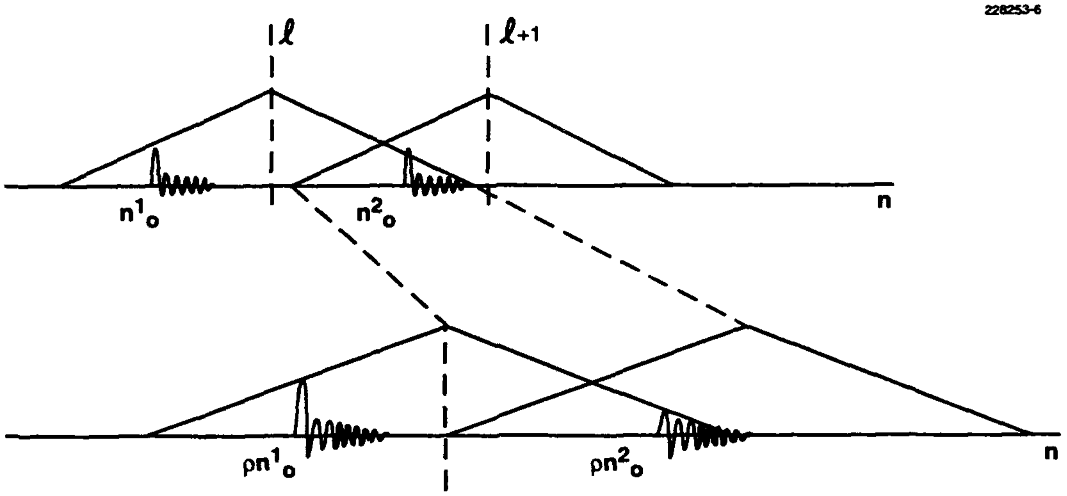

Figure 6. Short-time processing of long waveforms. The second event within frame ℓ is constrained to be the first event within frame $\ell + 1$.

Figure 7 illustrates application of the filter bank/overlap-add technique to a sequence of closely spaced and overlapping synthetic transients. Time-scale expansion has resulted in improved object audibility for synthetic signals of this kind (sums of damped sine waves) as well as for sequences of actual complex acoustic signals such as the sounds of a closing stapler, a bouncing wrench, and percussion transients. An example of time-scale modification of a closing stapler is shown in Figure 8. Figures 7 and 8 demonstrate excellent time resolution and spectral fidelity in the time-scaled reconstruction. In both cases, the analysis window is a 20-ms triangle.

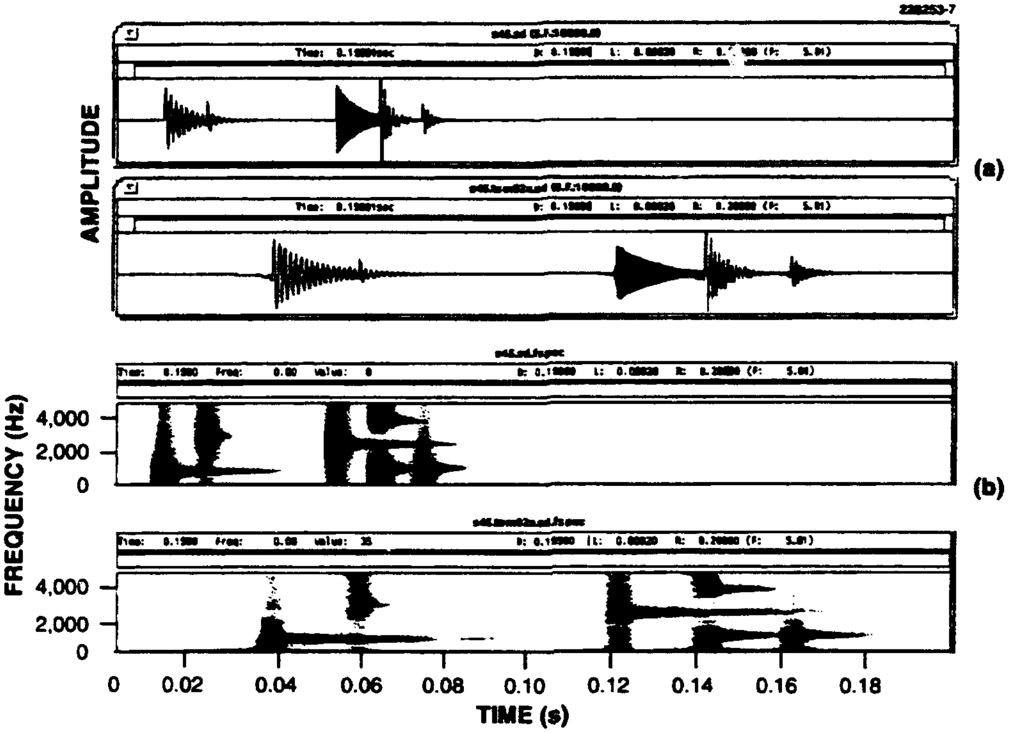

Figure 7. Time-scale expansion (×2) of sequence of transients using filter bank/overlap-add: (a) original and time-expanded waveform and (b) spectrograms of (a).

3.4 Discussion

The subband approach described in this section has been demonstrated to be effective in modifying sums of sequential, rapidly damped sine waves. In processing steady tones, however, the reconstruction has been found to add an unwanted amplitude modulation to the temporal envelope. This distortion is due likely to an inability of the system to synchronize over adjacent frames when no strong event is present. Furthermore, in processing complex signals in a noise background the system introduces distortion in the form of gurgles. Sections 4 and 5 address these problems.

Some fundamental questions also remain, the most basic of which involves the condition that in time-scale modification the time-scaled envelope should be preserved through appropriate phase control; other options may be considered. For example, in Figure 4 the two sequential events become more audible with or without phase correction. Nevertheless, the quality is distinctly different; without phase correction the reconstruction can exhibit a tinniness ("choral effect") from dispersion as well as a dullness in the attacks. Currently, the perceptual consequences of phase in complex acoustic signals are being investigated. Related to this question is the ability to show,

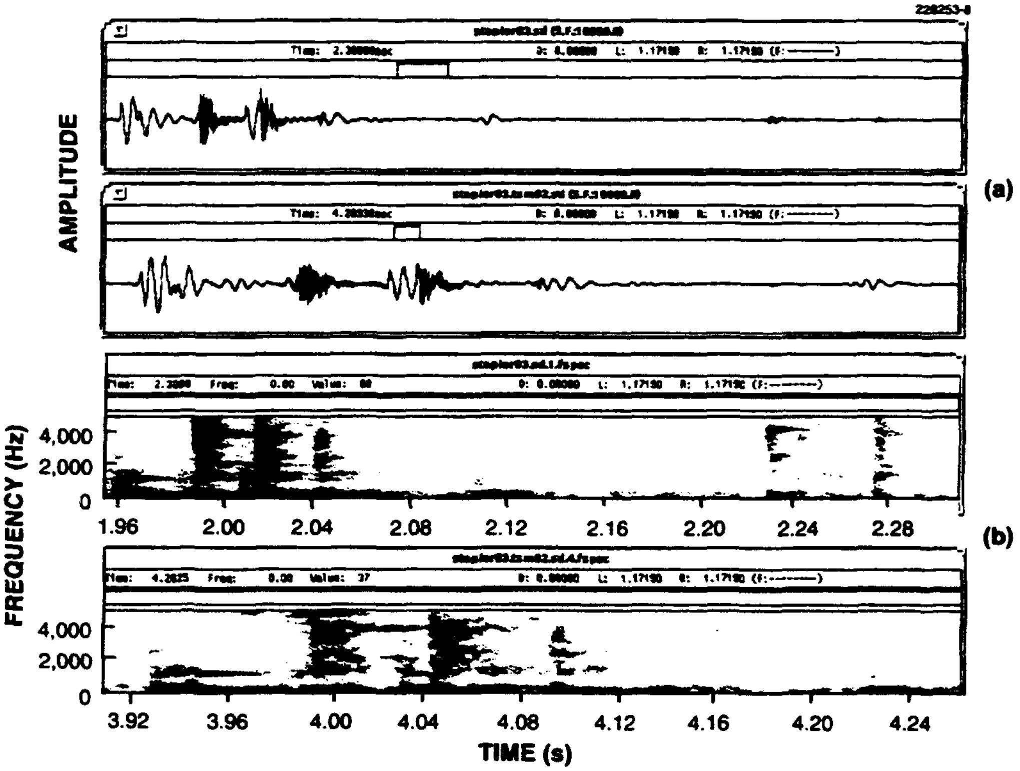

Figure 8. Time-scale expansion (×2) of a closing stapler using filter bank/overlap-add: (a) original and time-expanded waveform and (b) spectrograms of (a).

analytically, how the technique can achieve an approximate desired envelope by constraining sine-wave phase relations at only a few sample locations; phase relations resulting at other locations are not predictable.

Another issue involves how one might account for more than two events under a single analysis window and how subband components should be distributed among these events. In this case, the more similar the frequency content of the events, the more difficult it becomes to maintain their temporal structure by subband clustering. A rigorous development of these constraints and trade-offs is pending. Alternative methods of combining overlapping frames are also being considered. For example, adding (or interpolating) unwrapped phase functions of a single channel over adjacent frames is being explored as a procedure for adding overlapping time waveforms. Such an approach may lead to a graceful blending of the sine-wave and subband methods of time-scale modification.

4. BACKGROUND PRESERVATION

For a specific class of signals, the subband modification system satisfies the first condition for high-fidelity time-scale modification (i.e., approximate preservation of the time-scaled temporal envelope), but a problem arises in preserving the background. The problem, which is due likely to unwanted frame-to-frame correlation, is that this method introduces background artifacts. This section presents a technique to preserve the background character. This approach is developed in a subband framework and is integrated with the current method for object modification.

4.1 Fullband Spectral Shaping

One approach to synthesizing and modifying the background $b(n)$ is to spectrally shape white noise with a linear filter having a system function that matches the background spectral density. Time-scale expansion can be performed by passing noise through the filter for an extended period of time. The approach first requires estimating the background spectral density by averaging spectra of the observed signal $s(n)$. Let the short-time Fourier transform magnitude (STFTM) of $s(n)$ be given by

$$|S(\omega; mL)| = |\sum_n w(n - mL)s(n)\exp[jn\omega]| \quad , \qquad (17)$$

where L is the frame interval and $w(n)$ is the sliding analysis window. Assuming for the moment that the object $x(n)$ is not present, i.e., $s(n) = x(n) + b(n) = b(n)$, an estimate of the spectral density of the background $b(n)$ on the mth frame is obtained by averaging the squared STFTM (i.e., averaging the periodogram):

$$B(\omega; mL) = \alpha|S(\omega, mL)|^2 + (1 - \alpha)B(\omega; (m-1)L) \quad . \qquad (18)$$

This method is similar to the Welch method of spectral estimation [11].[6] When the background is a stationary random process, it can be shown that the expected value of $B(\omega; mL)$ is a smooth v϶rsion of the desired density

$$E[B(\omega; mL)] = \beta \int_0^{2\pi} B_i(\tau)W(\omega - \tau)d\tau \quad , \qquad (19)$$

[6] In addition, some mild smoothing can be performed across frequency via windowing the autocorrelation function associated with $B(\omega; mL)$ [11].

where $B_i(\omega)$ is the underlying spectral density of the background, $W(\omega)$ is the Fourier transform of the window $w(n)$, and β is a function of the window length. However, because an object generally may be present, the averaging operation in Equation (18) is performed only when the mth frame contains all background. The mechanism for this decision relies on a subband representation of the background and is given in Section 4.2.

A time-varying impulse response of a linear system can be associated with $B(\omega, mL)$ and is given by the inverse Fourier transform of $B(\omega; mL)^{1/2}$,

$$h(n; mL) = 1/2\pi \int_0^{2\pi} B(\omega; mL)^{1/2} \exp[jn\omega] d\omega \quad , \tag{20}$$

which is a zero-phase (symmetric) response. The simulated background over the mth frame is then given by

$$\hat{b}(n; mL) = h(n; mL) * [w(n - mL)e(n)] \quad , \tag{21}$$

where $e(n)$, a white-noise input, is multiplied by the sliding analysis window. Because the window $w(n)$ and frame interval L are designed so that $\sum_m w(n - mL) = 1$ (i.e., the overlapping windows form an identity), the overlapping sequences $\hat{b}(n; mL)$ can be summed to form the synthesized background

$$\hat{b}(n) = \sum_m b(n; mL) \quad . \tag{22}$$

When the background is stationary, the underlying impulse response is fixed so that as m becomes large, $h(n; mL)$ is approximately a time-invariant response $h(n)$. For large n, therefore,

$$\hat{b}(n) \approx h(n) * e(n) \quad , \tag{23}$$

and thus the background is approximately the output of a time-invariant linear filter.

To perform time-scale modification, a new window $w'(n)$ and frame interval L' are selected such that $\sum_m w'(n - mL') = 1$, and the factor L'/L equals the desired rate change factor ρ, which is assumed rational. The resulting time-scaled waveform is

$$\tilde{b}(n) = \sum_m h(n; mL') * [w'(n - mL')e'(n)] \quad , \tag{24}$$

where $e'(n)$ is the white-noise input generated on the new time scale. As in the baseline system, when the background response is stationary, for large n the synthesized background approaches the output of a fixed linear filter

$$\tilde{b}(n) \approx h(n) * e'(n) \quad , \tag{25}$$

where $h(n)$ is the time-invariant impulse response.[7]

4.2 Object/Background Assignment

To discriminate between object and background, a subband representation is used with bandpass filters $h_k(n)$, designed as described earlier to form a perfect reconstruction filter bank. Energy changes within each band are used to determine those regions of the spectrum that contain objects and those that contain background.

In making the object/background subband assignment, the instantaneous energy of the output of the kth channel for the mth frame is defined as

$$e_k(m) = \sum_n |y_k(n; mL)|^2 \quad , \tag{26}$$

where the complex filter output $y_k(n)$ is the convolution of the kth bandpass filter with the windowed input sequence

$$y_k(n; mL) = [w(n - mL)x(n)] * h_k(n) \quad . \tag{27}$$

The average energy of each filter output can then be computed as

$$\hat{e}_k(m) = \beta e_k(m) + (1 - \beta)\hat{e}_k(m - 1) \quad , \tag{28}$$

and the average variance of the instantaneous energy about its mean level is obtained as

$$\hat{\sigma}_k^2(m) = \beta[e_k(m) - \hat{e}_k(m)]^2 + (1 - \beta)\hat{\sigma}_k^2(m - 1) \quad , \tag{29}$$

[7] This method is similar in style to the noise modification approach of Serra [18], which is described in Appendix A.

where the "hat" denotes an estimate of the variance of the energy associated with the underlying random process. Finally, a detection statistic is defined as

$$D_k(m) = [e_k(m) - \hat{e}_k(m)]^2/\hat{\sigma}_k^2(m) \quad , \tag{30}$$

and a detection algorithm is formed as

$$D_k(m) > T - -- > \text{Object Band} \tag{31a}$$

$$D_k(m) \leq T - -- > \text{Background Band} \quad , \tag{31b}$$

where T is a threshold level. In addition to being used for selecting bands for particular modification schemes, the detection statistic also controls the averaging of the background spectrum because averaging the instantaneous spectrum in Equation (18) occurs only when all channels are declared background.

Under the condition that the energy function $e_k(m)$ is represented by a Gaussian random process, it is possible to formalize the detection algorithm. Specifically, it can be shown that the detection algorithm in Equation (31) follows from a significance test where the probability of false alarm is a function of the detection threshold T [19,20]. The band-dependent nature of the detection statistic Equation (30) reflects the constant false alarm rate across bands for a specific selection of T.

4.3 A Composite Subband System

Because the object and background modification have both been derived in the context of a subband representation, it is straightforward to merge the two algorithms into a single system.

4.3.1 Modification

Following the object/background subband assignment, a composite background filter is formed as

$$i_b(n; mL) = \sum_k h_k(n) \text{ for all k with } D_k(m) \leq T \quad , \tag{32}$$

and a composite object filter is formed as

$$i_x(n; mL) = \sum_k h_k(n) \text{ for all k with } D_k(m) > T \quad . \tag{33}$$

The time-varying background filter is applied to the average background spectral estimate to form a synthesis filter for the background. This filter is given in the frequency domain as

$$H_b(\omega; mL) = I_b(\omega; mL)B(\omega; mL) \quad . \tag{34}$$

For each segment the background synthesis filter is then driven by windowed white noise and the synthesized segments are summed to form the time-scaled background process

$$\tilde{b}(n) = \sum_m h_b(n; mL') * [w'(n - mL')e_b'(n)] \quad , \tag{35}$$

which is a short-time Fourier transform overlap-add realization. The subscript b is added to the white noise to distinguish it from a second noise process later to be used for object modification.

Conceptually, one can think of the object filter as providing channels that are to be phase-synchronized as in Section 3.2. The modified channel signal becomes

$$\tilde{y}_k^p(n) = \tilde{a}_k^p(n)\cos[\rho\tilde{\theta}_k^p(n) + \phi_k^p] \quad , \tag{36a}$$

where

$$\phi_k^p = \theta_k^p(n_o^p) - \rho\tilde{\theta}_k^p(\rho n_o^p) \quad . \tag{36b}$$

The superscript p refers to the first or second subband cluster (see Section 3.2) and the subscript k now refers to only those subbands designated as "object." The object component is then given by

$$\tilde{x}(n) = \sum_p \sum_k \tilde{a}_k^p(n)\cos[\rho\tilde{\theta}_k^p(n) + \phi_k^p] \quad , \tag{37}$$

where the sine waves are summed for each cluster.

In merging the two filter bank modification structures, one for object and one for background, the system illustrated in Figure 9 results. The windowed input segment is decomposed using the filter bank structure of Section 2, and the subbands are labeled "object" or "background." The object components are phase-synchronized and modified according to subband clustering of Equation (37), while the background components are modified by synthesizing the time-expanded signal as in Equation (35). The time-expanded background and object components are summed to form the new signal

$$\tilde{s}(n) = \tilde{x}(n) + \tilde{b}(n) \quad , \tag{38}$$

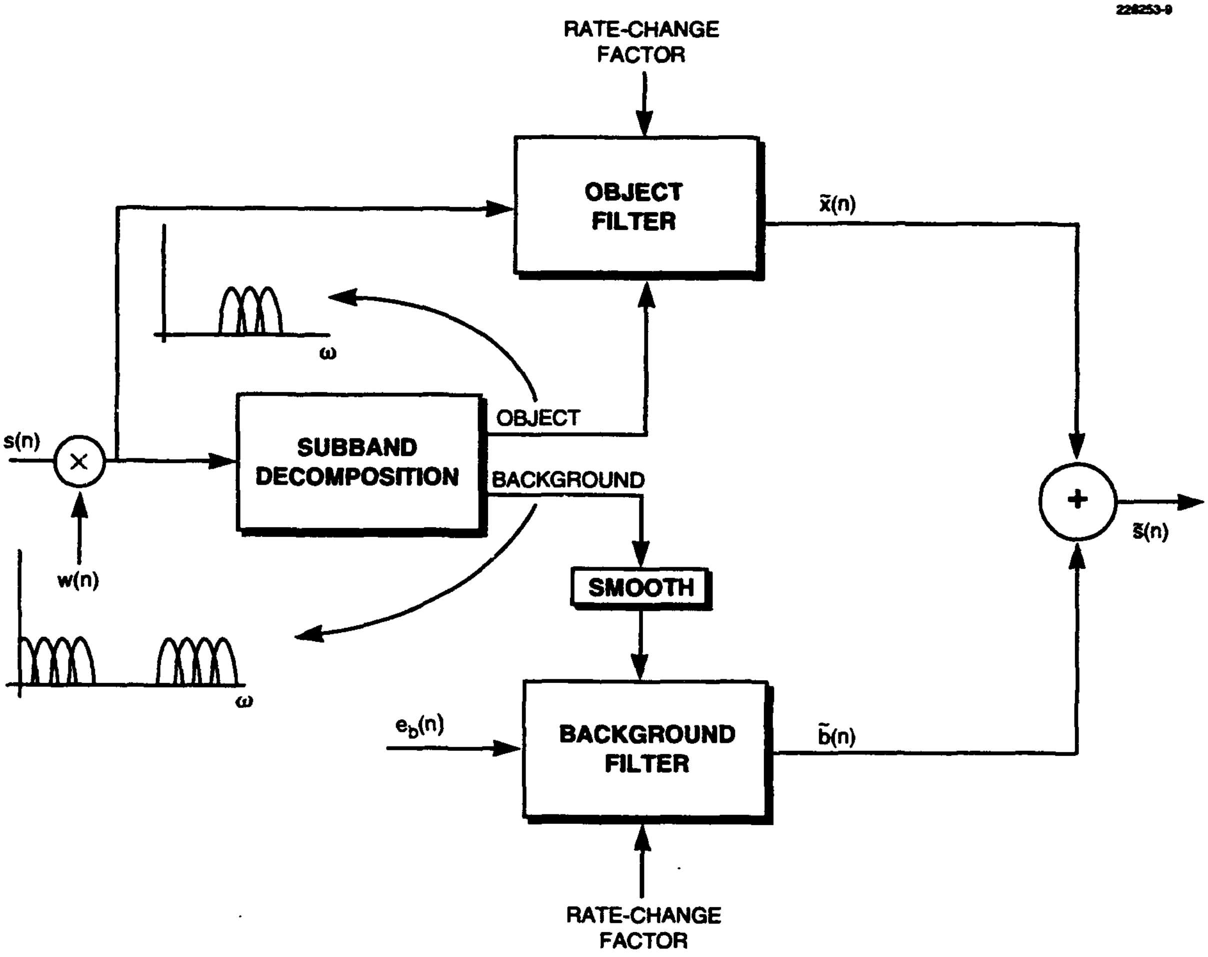

Figure 9. Composite subband time-scale modification algorithm.

which represents a time-scale expanded version of the original signal $s(n) = x(n) + b(n)$.

4.3.2 Examples

An example of the performance of the composite system is shown in Figure 10. The synthetic signal consists of white (uniform) noise added to three impulses that are perceived as "clicks." This signal is time-scale expanded by a factor of 2. The audibility of the closely spaced clicks is improved, and the background character is preserved. A second example is shown in Figure 11, where the signal is that of Figure 7 with an added background of white (Gaussian) noise; also illustrated is the instantaneous energy from the bandpass filter centered at 1,000 Hz.

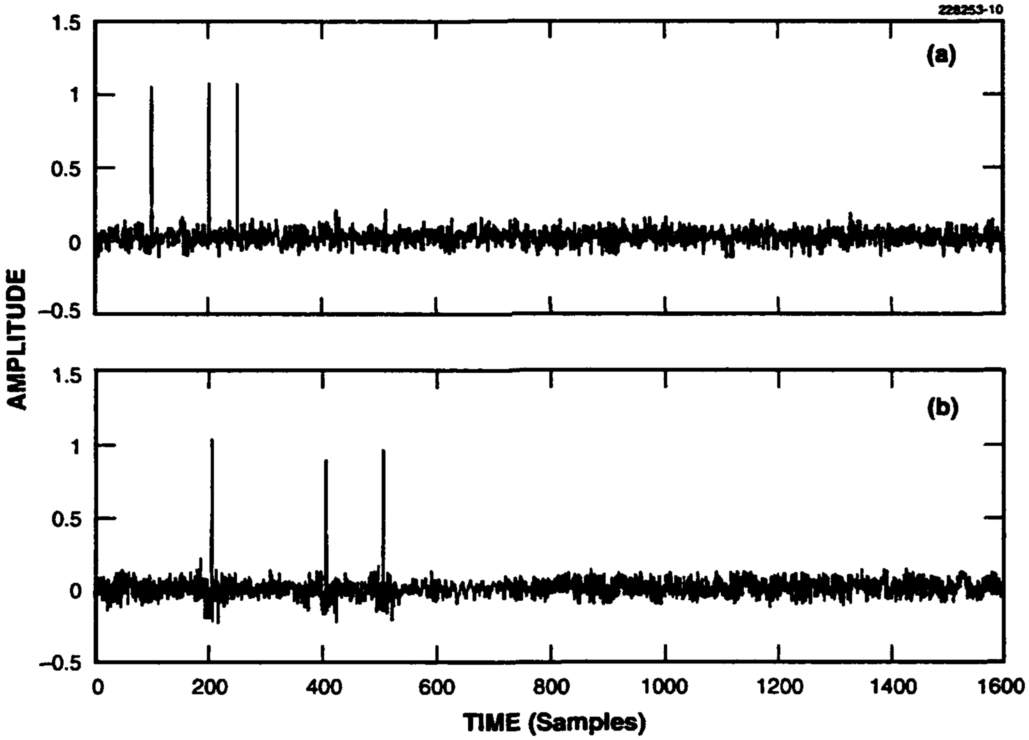

Figure 10. Time-scaled modification of click train in noise using composite subband system: (a) original and (b) time-expanded (×2).

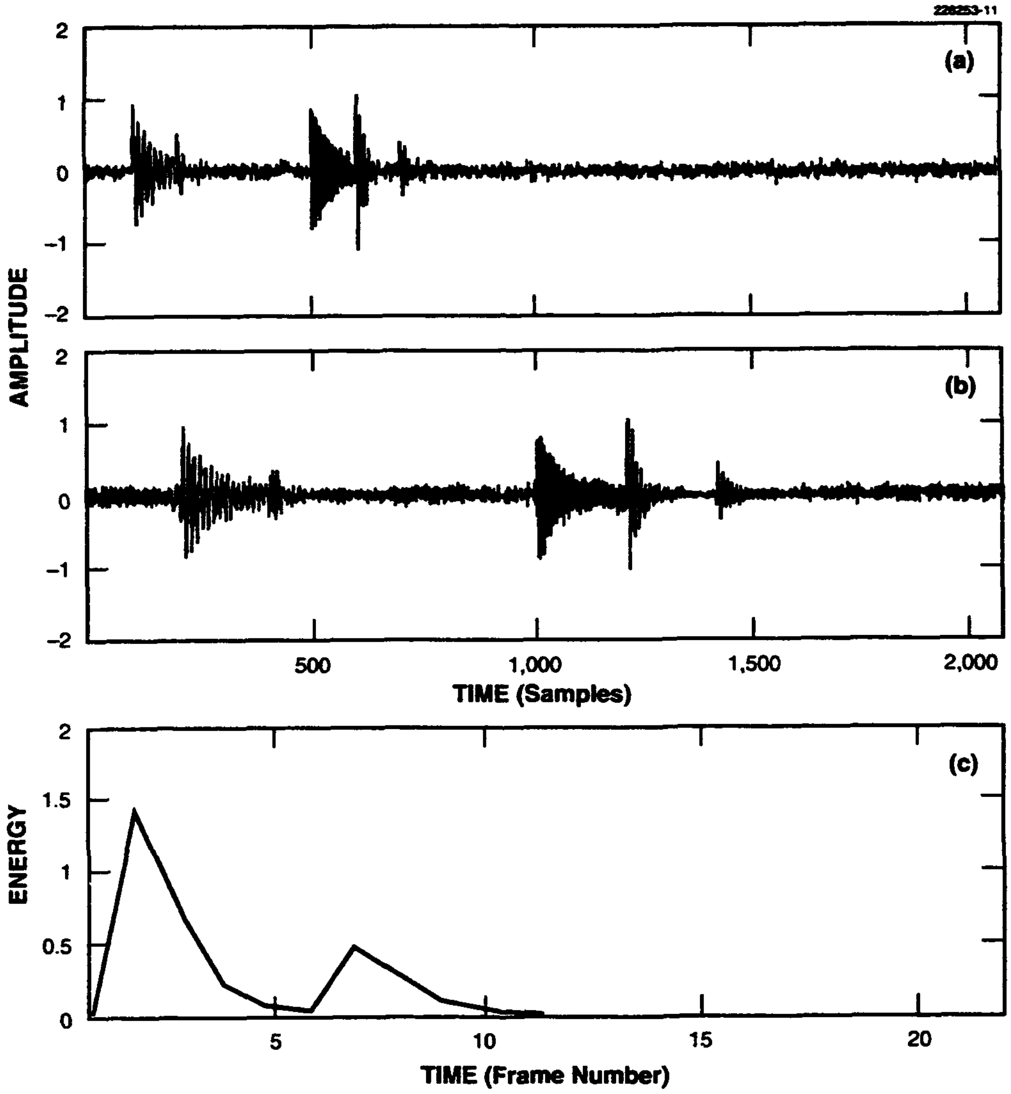

Figure 11. Time-scaled modificiation of complex signal in noise using composite subband system: (a) original, (b) time-expanded (×2), and (c) instantaneous energy from bandpass filter centered at 1,000 Hz.

4.4 Discussion

The background of the resulting modified waveform has the desired spectral shaping; however, provision has not been made for the temporal envelope of the background. Although this temporal structure will be reflected partly in energy variations through the spectrum, some fine temporal structure will not be preserved because phase of the original background signal is altered. For example, consider rain falling on a rooftop or water splashing against a buoy. In these cases, the fine structure of the temporal envelope that helps to characterize the sound may be lost by white-noise driven synthesis.[8]

One approach to recover temporal structure is to impose a time-scaled version of a temporal envelope estimated from the original background. This style is similar to the fullband temporal envelope invariance method described in Appendix B and resembles an approach recently (independently) proposed for time-scale modification of unvoiced speech [21]. An alternate approach to this problem is to apply a different input function, e.g., a Poisson process in the case of falling raindrops, rather than a uniform or Gaussian white-noise process. The background reproduction might also be improved by synthesizing background as well as object, in those subbands that were declared object bands. An alternate approach is to subtract an estimated object component from the input prior to background generation, and then declare the residual as the background to be modified over the full frequency band. This approach was used in a sine-wave context by Serra [18] and is discussed in Appendix A.

Another consideration involves selection of the subband filter width for detection and synthesis. Wide filters appear to be preferred for click detection, while narrow filters are preferred for tone detection; likewise for synthesis. A narrowband subband analysis will likely be best for both detection and synthesis because, when appropriate, narrowband filters can be summed to form wideband filters.

[8] If, however, the background is perceived as characteristic of the environment, this change may not be important to the listener.

5. STRAWMAN DESIGN FOR A GENERAL OBJECT CLASS

In the development of the modification system in Figure 9, it was implicit that an object consists of sums of rapidly damped sine waves with distinct occurrence times. In practice, objects can be of many varieties, including long tones and tones plus clicks; they may be partly or even fully stochastic, arising from turbulence in the signal generation process. Examples of such signals abound in underwater, biologic, and music sounds. To account for this general signal class, this section develops a semiautomated (strawman) design that is tested on a class of underwater sounds under a variety of signal-to-noise conditions.

5.1 Design

The strawman design is illustrated in Figure 12. In contrast to the earlier system, this design assumes two object classes, tones and nontones, and the method of modification depends on the class. It is assumed that a signal contains objects of only one class and that the signal class is known a priori.

When the object is nontonal, it is time-scale expanded using the approach that was developed for background modification, i.e., spectrally shaping a white-noise process. However, rather than deriving the shaping filter from a slowly varying background spectral estimate, it is derived from the instantaneous STFTM, $|S(\omega; mL)|$, which is weighted by the object filter to form the shaping filter

$$H_x(\omega; mL) = I_x(\omega; mL)|S(\omega; mL)| \quad , \tag{39}$$

and the time-scaled object is obtained by exciting this filter with windowed white noise and summing the filtered segments

$$\tilde{x}(n) = \sum_m h_x(n; mL') * [w'(n - mL')e'_x(n)] \quad , \tag{40}$$

where the primes refer to the time-scale modified parameters and signals. These operations are similar to those used in the background modification process of Equation (35). The difference is that $e'_x(n)$ denotes a second noise process distinct from $e'_b(n)$, and the filter represents the object spectral characteristics and not those of the background.

When the object consists of tones, sine-wave analysis/synthesis is used to time-scale the object [6,9]. A sine-wave decomposition is first performed through peak-picking the short-time Fourier transform of $y(n)$. This procedure yields a set of complex amplitudes of the form $x_l(mL) = a_l(mL)\exp[j\theta_l(mL)]$, which are weighted by the object filter in Equation (33):

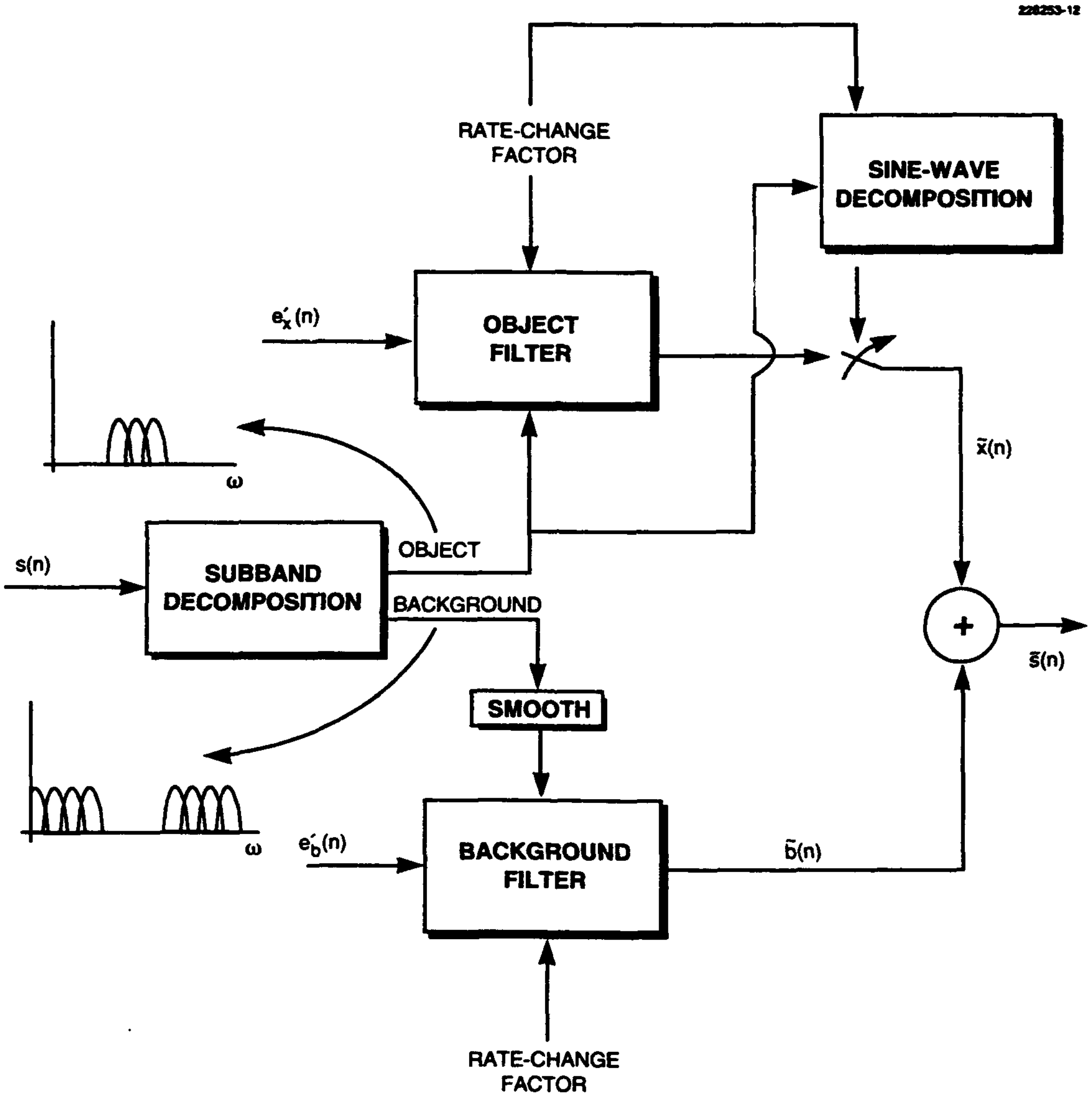

Figure 12. Strawman design for time-scale modification.

$$\tilde{x}_l(mL) = I_x(\omega_l; mL) a_l(mL) \exp[j\theta_l(mL)] \quad . \tag{41}$$

The weighted components are time-scaled as in the baseline sine-wave modification system of Appendix A, and the results are summed to form the modified object component $\tilde{x}(n)$.

The resulting time-scaled signal is given by

$$\tilde{s}(n) = \tilde{x}(n) + \tilde{b}(n) \quad , \tag{42}$$

where the object component $\tilde{x}(n)$ is generated as either modified tones or nontones. An automated procedure for selecting tones or nontones is now being developed (see Appèndix A). A feature of this approach is its fault tolerance to false alarms. If the detector declares object when background is present, then the (correct) average background spectrum is replaced by either the instantaneous spectrum (in the case of nontones) or samples of the instantaneous spectrum (in the case of tones). The deleterious effect of using the instantaneous spectrum is therefore limited because it represents a noisy version of the background spectrum.

5.2 Examples of Synthetic Signals

In informal testing of the fault tolerance of the modification algorithm, two 10-s intervals of Gaussian noise (white and colored) were processed without the presence of an object. In each case, the modified waveform was aurally similar in character to the original background, except for one low-level "ting" (roughly a millisecond in duration) in the white-noise case. Only this single artifact was perceived, although more than one false alarm occurred. Nevertheless, the presence of this artifact indicates the need to fine tune the threshold level in Equation (31) to achieve an appropriate trade-off between detections and false alarms.

The performance of the approach is illustrated in Figure 13 using a synthetic signal consisting of two tones and three clicks with a colored noise background. The original and the processed signals, time-scaled by two and three, are shown in Figure 13(a) in the time domain; corresponding spectrograms are shown in Figure 13(b). In this example, the object was assumed (a priori) to be tonal, i.e., background was synthesized with a white-noise-driven linear filter of Equation (35), while the object was synthesized with the sine-wave synthesis of Equation (41). The duration of the analysis window was set at 10 ms with a 2-ms frame interval. Figure 13(c) shows the instantaneous energy output, Equation (26), of the subband filter at 1,000 Hz. In this example the object became more audible while the background maintained its essential character, and no aural artifacts were introduced into the background. The tones were accurately reconstructed, but the clicks became somewhat tonal. When the time-scale expansion was performed using a noise-generated instead of a sine-wave-generated object, the clicks better maintained their character, but the tones became noise-like.

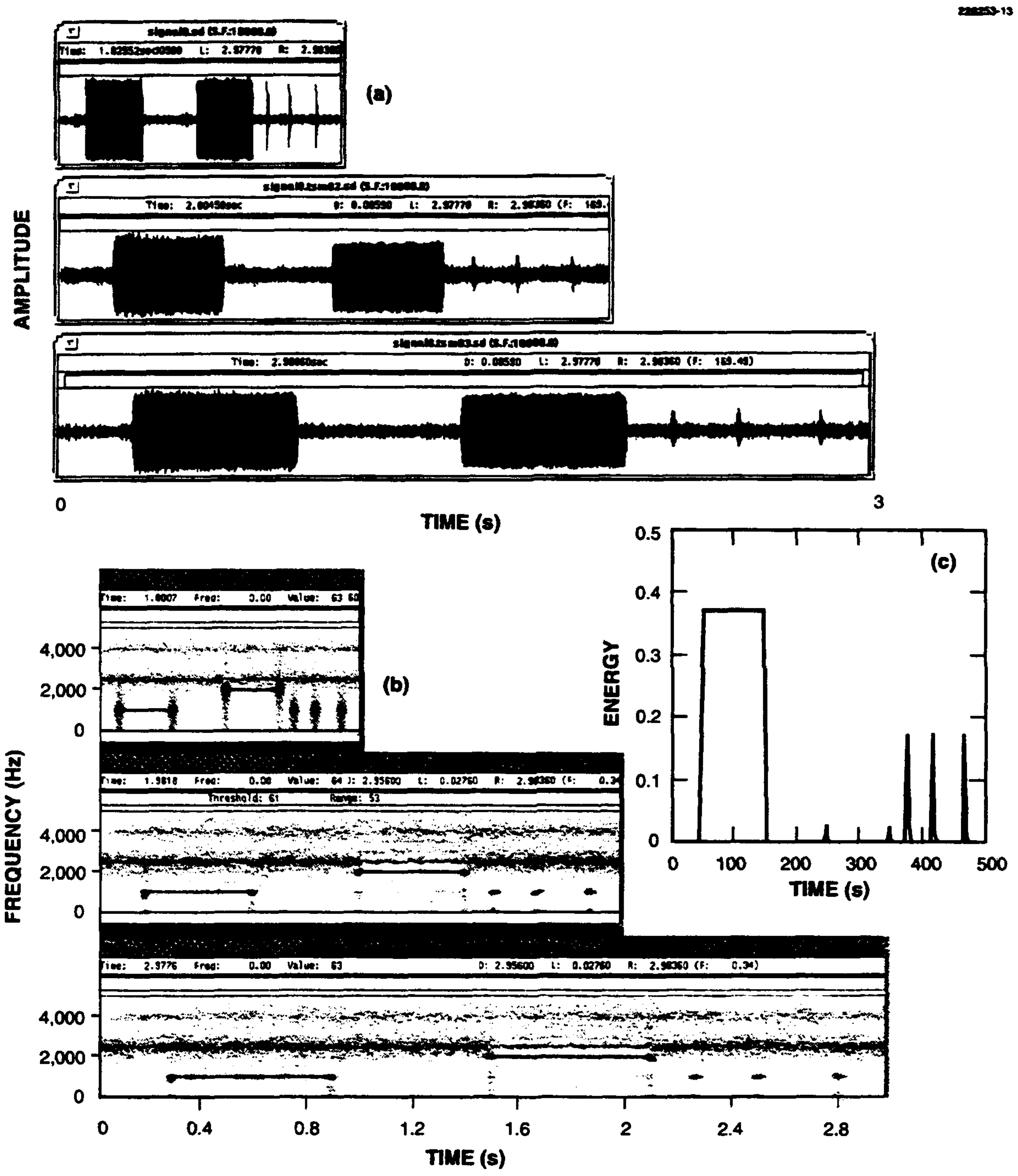

Figure 13. Time-scale modification of synthetic signal: (a) waveforms, (b) spectrograms, and (c) instantaneous energy from bandpass filter at 1,000 Hz.

5.3 Application to Underwater Acoustic Signals

Using the strawman design, a database of underwater sounds was time scaled by factors of 2, 4, and 8. The objects were manually classified as either tones or nontones and processed accordingly. The ocean backgrounds generally consisted of nonstationary noise. Six different signals were processed at three signal-to-noise ratios (SNRs)—high, medium, and low—and the results were judged by both experienced and inexperienced listeners. The listeners agreed that the object components were generally more audible, backgrounds remained ocean-like, and few artifacts were introduced. The latter result is somewhat surprising at low SNRs, where the detection algorithm, which is quite simple, was taxed. Processing was automated except for the tonal/nontonal object designation and the tailoring (to the degree of background stationarity) of certain time constants used in smoothing the spectral background and subband energy for detection. (They were not, however, changed across SNRs.)

Figure 14 shows an example of time-scaling (by two, four, and eight) an underwater acoustic signal in an ocean background at a high SNR. (In the modification, the synthesized backgound was reduced to make visible the object components that are not visible in the original.) In this case the object was synthesized with white-noise-driven subband filters. The analysis window length was 6 ms with a 2-ms frame interval. Figure 14 also shows the instantaneous and average energy functions used in object/background assignment from the output of the bandpass filter at 3,000 Hz. Figure 15 shows a different underwater signal that consists of low-pitch/low-frequency tones in ocean background. Because the object consisted primarily of tones, sine-wave synthesis was chosen for the object and a 10-ms analysis window was used. In both examples the object signal became more audible while the background maintained its essential character with the perception of few false alarms.

5.4 Discussion

These examples illustrate the potential of the strawman design to improve audibility but also indicate a number of problems. Objects with sharp attacks and very short duration can be smeared and distorted when generated by white-noise-driven synthesis. With large time-scale factors (e.g., a factor of 8), the object suffers from what is referred to here as the "bean bag effect," e.g., a hammer slamming against a metal plate is perceived as a bean bag falling on a wooden floor. A second problem is that deterministic backgrounds may lose their character. Some examples (splashing water, falling rain) were discussed in Section 4.4 along with possible solutions.

In the later problem, when it is known a priori that the object consists of rapidly damped sine waves, the composite subband system that was developed in Section 3 should be applied. The phase synchronization will better capture the temporal structure of the signal. Using such a segment from an underwater sound, Figure 16 compares the two modification systems (Figures 9 and 12). In addition to being visually sharper, the composite subband system with phase synchronization yields a reconstruction that is also audibly sharper.

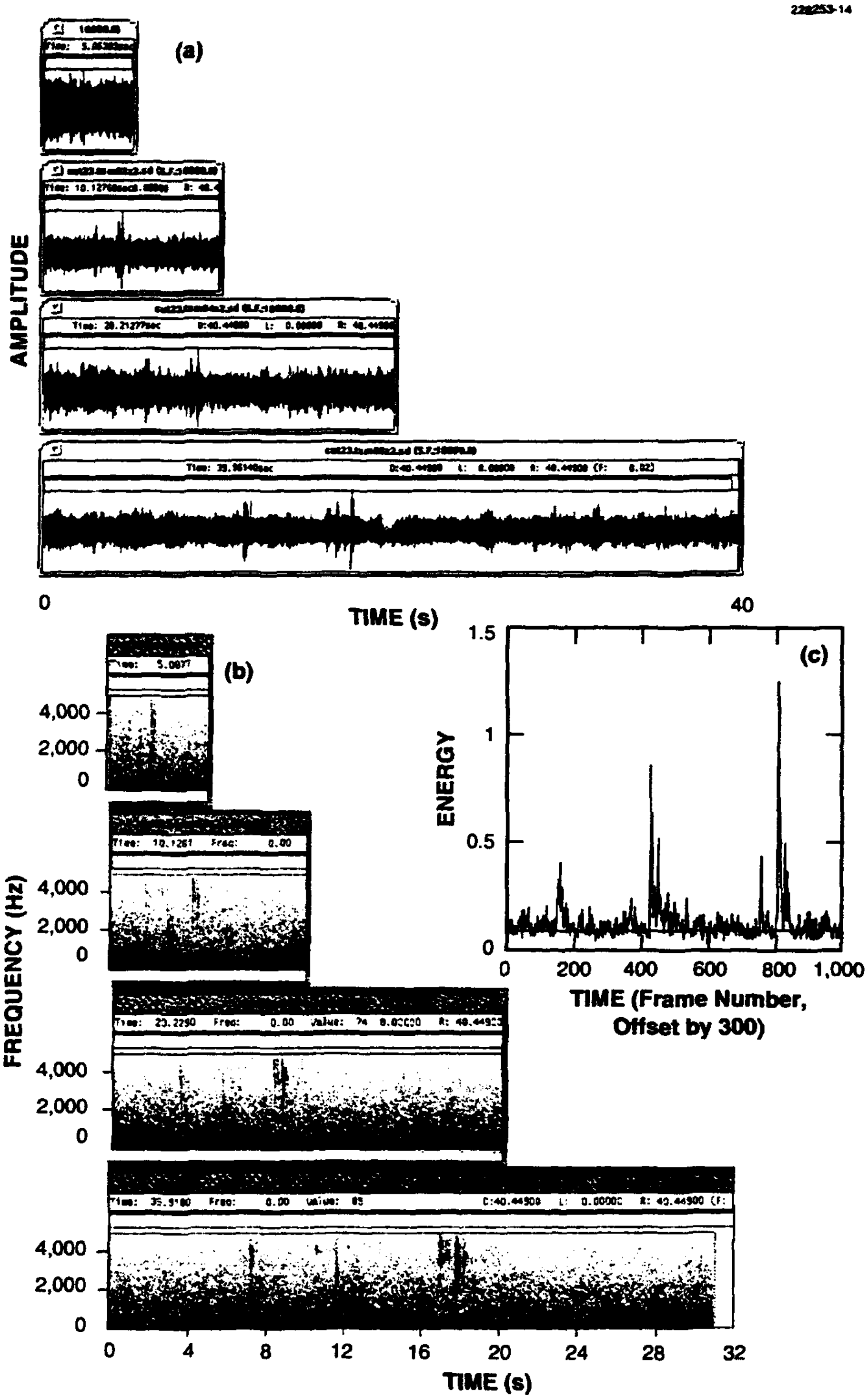

Figure 14. Time-scale modification of underwater signals using strawman design: (a) waveforms, (b) spectrograms, and (c) energy measures from bandpass filter at 3,500 Hz.

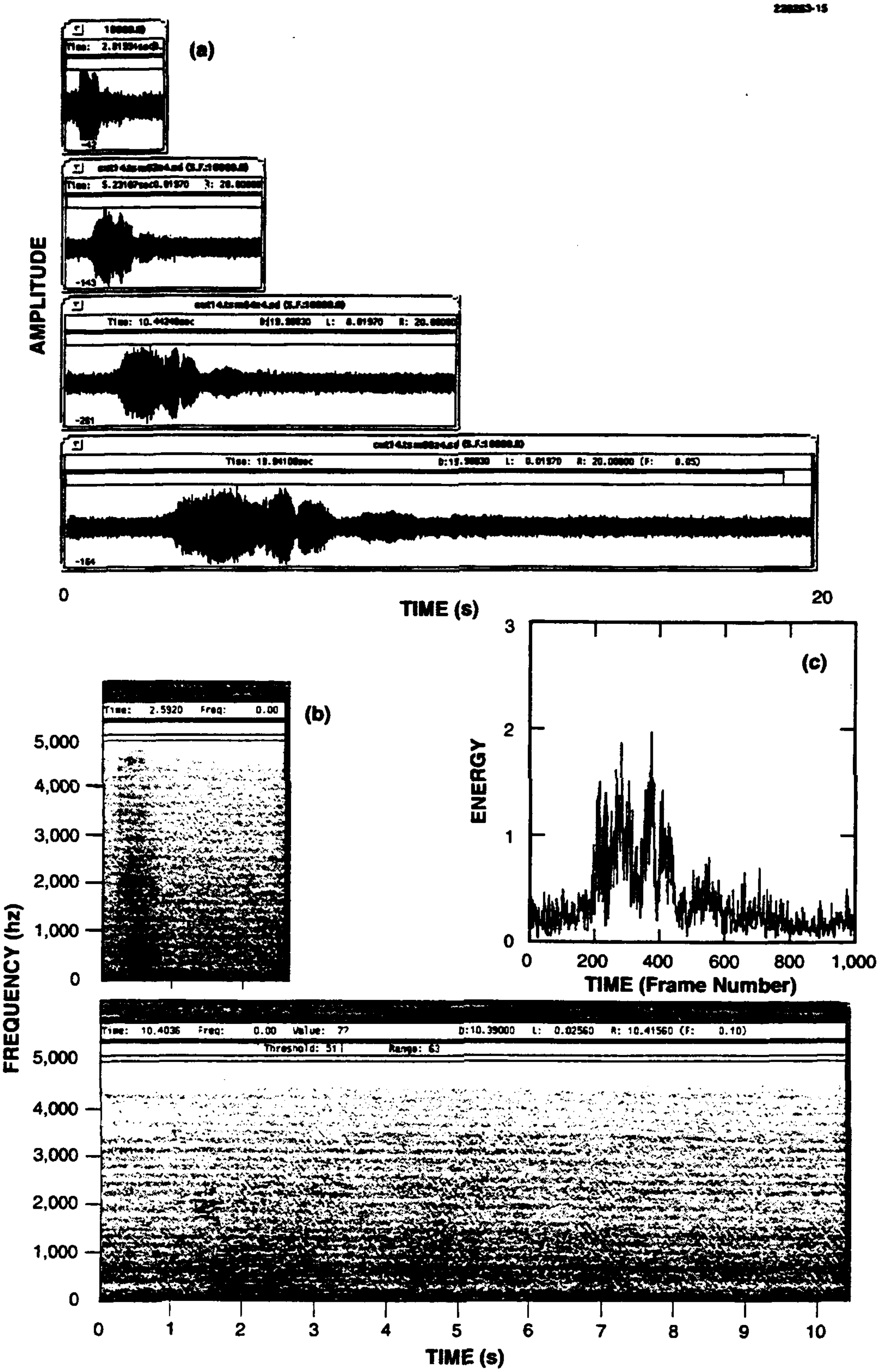

Figure 15. Time-scale modification of underwater signals using strawman design: (a) waveforms, (b) spectrograms (original, ×4), and (c) energy measures from bandpass filter at 500 Hz.

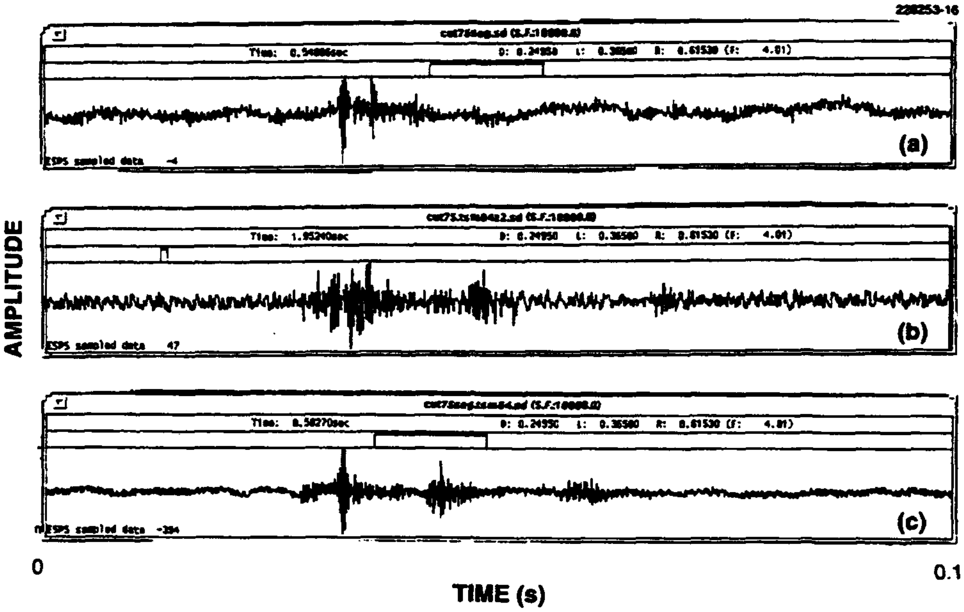

Figure 16. Comparison of modification algorithms: (a) original, (b) expanded (×4) with white-noise-driven synthesis, and (c) expanded (×4) with phase synchronization.

Another issue involves using two separate white-noise processes in Equations (35) and (40) to account for turbulence as a random process, distinct from a random background (a phenomenon that occurs often, for example, in underwater and music signals). Using the same white-noise process to synthesize both turbulence and background will likely make the resulting signal components less aurally distinct than when created from two independent random processes. This perceptual separation of sound is consistent with Bregman's paradigm of "sound segregation" requiring distinct sources [22]. The importance of the use of two different inputs, nevertheless, has yet to be clearly justified in the current application.

6. SUMMARY AND FUTURE WORK

A new subband approach was introduced for time-scale modification of complex acoustic signals consisting of sums of rapidly damped sine waves. The approach requires control of the phase of clusters of filter bank outputs and preserves the structure of the time-scaled temporal envelope of a signal while maintaining its spectral character. Appropriate spectral shaping and correlation in the time-scaled background is provided by controlling the subband amplitudes and filter bank inputs. The subband phase and amplitude and filter bank input control are derived from locations of object events in time and frequency. A frame-based generalization of the method gives phase consistency and background noise continuity across consecutive synthesis frames. A strawman design of a modification system was developed to account for a more general class of objects. Although significant audibility gains were achieved, much remains to be accomplished. Some of the important directions were discussed throughout the report; a summary of future efforts follows.

6.1 Object Synthesis

Three methods of object time-scale modification were described in this report: subband filtering with phase synchronization (for rapidly damped sine waves), sine-wave synthesis (for tonal signals), and white-noise-driven filtering (for random signals). A greater variety of signal states would be ideal, e.g., tones, clicks, damped sine waves, turbulence, and any combination of such states. Discrimination among these states is necessary to provide control for signal synthesis and time-scale modification. The most difficult and perhaps the most fascinating problem will be to distinguish background from turbulence. The notion of chaos may be useful in characterizing turbulence and hence in providing a means of making this discrimination [23]. Refined methods of time-scale modification for each signal state will require maintaining the distinctive quality of the object component, as well as its perceived separation from its background and neighboring object components.

6.2 Background Synthesis

The considerations for background modification are similar to those for the object component. The character of the background is more or less adequately reproduced, depending on its complexity; driving a linear system with white noise will not necessarily reproduce the shape of the time envelope. Alternative synthesis schemes are to use a different driving function (e.g., a Poisson process), to shape the modified background with a time-scaled temporal envelope, and to use harmonic sine-wave phase dithering.

6.3 Detection

With a primitive detector based on subband energy, the results were surprisingly good in distinguishing object from background. Nevertheless, one of the more difficult problems is setting time constants and smoothing parameters within the detection algorithm. Estimation and application of these parameters should be made as adaptive and dynamic as possible. The detector should look across subband filter outputs as well as time; more generally, it may be desirable to merge detection with object classification schemes that capitalize on both past and future analysis frames.

6.4 Multiresolution Representation

Time-scale modification is likely to benefit from multiresolution analysis/synthesis. Some signals consist, for example, of both low-frequency tones and high-frequency clicks, requiring analysis with fine frequency resolution for low-frequency components and fine time resolution for high-frequency components. Moreover, multiresolution synthesis can be essential in time-scale expansion of such signals. For example, when sine-wave synthesizing a low-frequency tone, specification of phase boundary conditions may lead to a constraint inconsistent with the sine-wave frequency, thus requiring a long synthesis interval relative to that for a high-frequency tone. Finally, multiresolution analysis/synthesis may also be useful in eliminating the need for preserving a fullband temporal envelope. Under the assumption that the auditory system uses logarithmically spaced critical bands in its analysis, independently preserving the temporal envelope in each band may be sufficient for perception; this is an easier task than maintaining the fullband envelope.

APPENDIX A
Signal Modification by Sine-Wave Analysis/Synthesis

This appendix describes time-scale modification by sine-wave analysis/synthesis [6,7]. A number of new methods are introduced to preserve a noise-like background.

A.1 Sine-Wave Representation of Acoustic Signals

The sine-wave representation of a signal is given by a sum of sine waves with time-varying amplitudes, frequencies, and phases [7]:

$$s(t) = \sum_{k=1}^{P} A_k(t) \ \cos[\Theta_k(t)] \quad . \tag{A.1}$$

The amplitudes and phases for the kth sine wave are denoted by $A_k(t)$ and $\Theta_k(t)$, respectively, and the time-varying frequency of each sine wave is given by the derivative of the phase. This frequency is denoted by $\omega_k(t) = \dot{\Theta}_k(t)$ and is sometimes referred to as the kth "frequency track." Although this model was originally formulated for speech signals, it is also capable of representing nonspeech signals such as underwater, biological, and music sounds, and signals consisting of two or more components (e.g., an underwater acoustic signal added to an interfering background).

A.1.1 Analysis/Synthesis

Using the sine-wave model Equation (A.1), an analysis/synthesis system has been developed [6,7]. Because measurements are made using digitized sounds, sampled-data notation is used. On each analysis frame the sine-wave parameters are estimated at time samples $n = mQ$, where the frame number $m = 0, 1, 2...$ and where Q is the number of samples in the frame interval. The dependence of the sine-wave parameters on the discrete time variable n is therefore replaced by their dependence on the frame number m, e.g., $A_k(n)$ is replaced by $A_k(mQ)$ or for simplicity by $A_k(m)$. A 5- to 10-ms frame interval has been found to produce high-quality reconstruction for most signals of interest. The analysis window (a Hamming window, typically 10 to 25 ms in duration), denoted by $w(n)$, is placed symmetrically relative to the origin, which is defined as the center of the current analysis frame. A discrete short-time Fourier transform (STFT) is then computed over this duration with a fast Fourier transform (FFT), typically, 1,024 or 2,048 points. The frequencies $\omega_k(m)$ are estimated by picking the peaks of the uniformly spaced FFT samples of the STFT magnitude. The sine-wave amplitudes $A_k(m)$ and phases $\Theta_k(m)$ at the center of each analysis frame are given by the amplitude and phase of the STFT at the measured frequencies.

The first step in synthesis requires association of the frequencies $\omega_k(m)$ measured on one frame with those obtained on a successive frame, which is accomplished with a nearest-neighbor matching algorithm that incorporates a birth-death process of the component sine waves, i.e., the sine waves are allowed to come and go in time. The amplitude $A_k(m)$ and the phase $\Theta_k(m)$ parameters are then interpolated from one frame to the next at the matched frequencies so that they may be upsampled to the original sampling rate. The amplitude is interpolated linearly and the phase is interpolated with a cubic polynomial, the latter being done using the methods described in Quatieri and McAuley [6] and McAulay and Quatieri [7]. The interpolated amplitude and phase components are then used to form an estimate of the waveform, according to Equation (A.1). Figure A-1 illustrates the entire analysis/synthesis process.

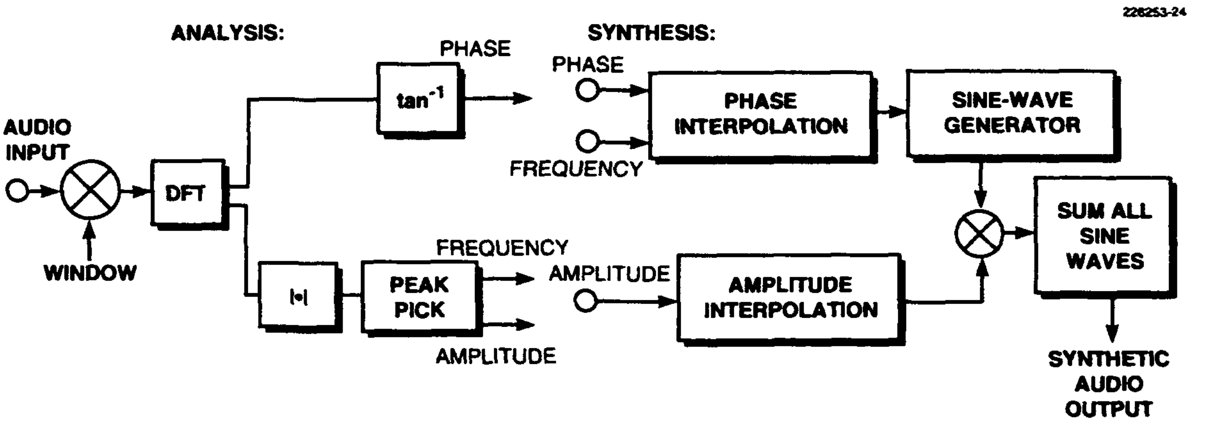

Figure A-1. Sine-wave analysis/synthesis.

A.1.2 Application to Complex Acoustic Signals

The enhancement problem is concerned with two signal classes: the desired complex acoustic signal $x(n)$, i.e., the object, and the background signal $b(n)$. These signals are added to form the measured signal $s(n) = x(n) + b(n)$. About 25 complex acoustic signals were collected from audio recordings, e.g., a bouncing can, a slamming book, and a closing stapler. These signals were selected to have sharp attacks with a variety of time envelopes and spectral resonances. The background signals are both synthetic and real, and include man-made signals (e.g., white noise), and natural backgrounds (e.g., ocean noise).

Although sine-wave analysis/synthesis is not strictly an identity, the sine-wave reconstruction of these acoustic signals and background were found to be nearly perceptually indistinguishable from the originals. To maintain the time and frequency resolution required to reconstruct these signals, the duration of the analysis window $w(n)$, frame interval L, and the number of sine-wave peaks P are adapted to the signal type. However, because the window duration was set to obtain adequate spectral resolution, some temporal smearing can occur for short signals and signals with sharp attacks and is perceived as a slight dulling of the sound.

The analysis/synthesis expresses signals in terms of a functional model describing the behavior of its sine-wave components. The sine-wave representation therefore provides a good framework for developing signal enhancement techniques based on transforming the functional descriptors.

A.2 Time-Scale Modification

In time-scale modification, the magnitude, frequency, and phase of the sine-wave components are modified to expand the time scale of a signal without changing its frequency characteristic.

A.2.1 Algorithm

Consider a time-scale expansion by a factor of β. By time-expanding the sine-wave frequency tracks, i.e., $\omega_k(\beta t) = \dot{\Theta}_k(\beta t)$, the instantaneous frequency locations and magnitudes are preserved while modifying their rate of change in time. Because $d/dt[\Theta_k(t\beta)/\beta] = \omega_k(\beta t)$, this modification can be represented by

$$\tilde{s}(t) = \sum_{k=1}^{N} A_k(\beta t) \ \cos[\Theta_k(\beta t)/\beta] \quad . \tag{A.2}$$

The discrete-time implementation of Equation A.2 requires mapping the synthesis frame duration Q to βQ, and then sampling the modified cubic phase and linear amplitude functions derived for each sine-wave component over this longer frame.

An example of slow-motion audio replay applied to the closing stapler of Figure 8 is illustrated in Figure A-2. Figure A-2(b) and (f) shows the high fidelity of the signal reconstruction without modification; while Figure A-2(c) and (g) gives the time expansion by two, both examples using a 7-ms analysis window and a 2-ms frame interval. Each frequency component lingers over a longer time duration than in the original, the effect of which is greater separability of the time events and a sharpening of the spectral resonances. In informal listening, the audibility of the stapler's rapidly changing sequence of events is enhanced. However, some of the fine temporal structure, e.g., the sharp attack times, has been smeared; in contrast, the subband approach illustrated in Figure 8 has better preserved this temporal structure and thus provides an aurally sharper reconstruction.

Attempting to improve the time resolution by shortening the time-domain analysis window $w(n)$ and decreasing the frame interval L can cause various forms of distortion. Poor spectral resolution from inadequate peak-picking may result from a short analysis window, while phase distortion from possibly inconsistent phase constraints in synthesis may result from a short frame interval. Figure A-2(d) and (h) shows an example where a 3-ms window (which is about the length of the prototype filter used in subband synthesis, thus giving the two methods roughly equal temporal resolution) and a 1-ms frame have led to both temporal and spectral distortion. Figure A-3 compares the sine-wave and subband approaches for a synthetic signal consisting of a series of damped sine waves. The sine-wave modification uses a 3-ms window and a 1-ms frame. Spectral

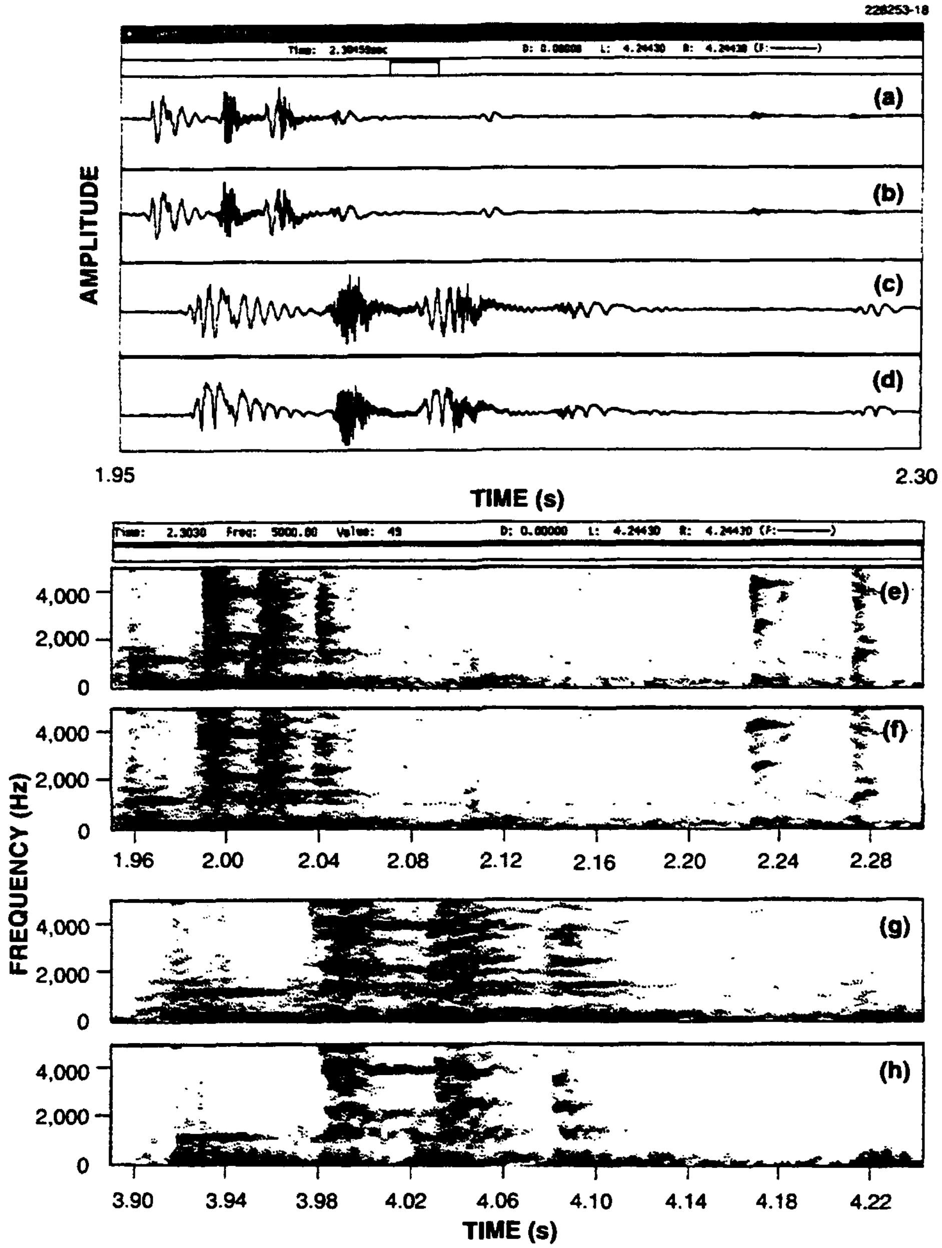

Figure A-2. Example of time-scaled expansion (×2) using sine-wave analysis/synthesis: (a) original, (b) reconstruction—7-ms window, 2-ms frame, (c) expanded—7-ms window, 2-ms frame, and (d) expanded—3-ms window 1-ms frame. [(e)–(h) are spectrograms of (a)–(d).]

distortion in the sine-wave modification is evident from the spectrogram displays, while a temporal smearing is seen in the sharp attacks of the waveform. In these cases, the distortions were perceived as a tinniness, due possibly to spectral distortion, and occasionally as a choral effect, perhaps from a dulling of attacks and phase distortion.

A.2.2 Problem of Preserving the Background

Signal modification should be performed so that the character of the modified background is similar to the original. For random backgrounds, e.g., ocean noise, time-scale expansion by factors of roughly 2 or more may result in synthesized sine waves that are perceived as tones, thus destroying the noise-like property. The problem is that long synthesis frames, resulting from time expansion, impose an unwanted time correlation on the sine-wave amplitudes and phases. To avoid this objectionable tonality, a method to decorrelate the sine-wave phases across successive frames is being developed.

The essence of the technique is to add a random sequence to each sine-wave phase after doing cubic phase interpolation in the synthesis stage. This perturbation, although decorrelating the background phases, will also decorrelate the phases of the object signal. Consequently, an adaptive procedure is being developed that adds the phase perturbation only when the object is not present. One possibility is to form a modified phase for each frame as

$$\tilde{\Theta}_k(n) = \Theta_k(n) + \epsilon_k(n) \quad , \tag{A.3a}$$

where

$$\epsilon_k(n) = \pi \delta D_k(n) \tag{A.3b}$$

with δ a random number falling uniformly in the interval $[-1, 1]$, and $D_k(m)$ takes on the value zero when an object is present for the kth sine wave and one otherwise. Detection is performed in uniformly spaced frequency bands by comparing the instantaneous energy in each band with a threshold derived from a running average energy; this approach is similar in style to the detection scheme in Section 4.2. The approach has shown promise in preserving the background noise character while keeping the desirable properties of the time-scaled object signal. However, a problem of overly whitening a colored background arises, because whitening the phase of a sine wave results in what is referred to as the "bean bag effect," e.g., a hammer slamming against a metal plate is perceived as a bean bag falling on a wooden floor. Two alternative methods that result in the proper temporal correlation are now briefly described.

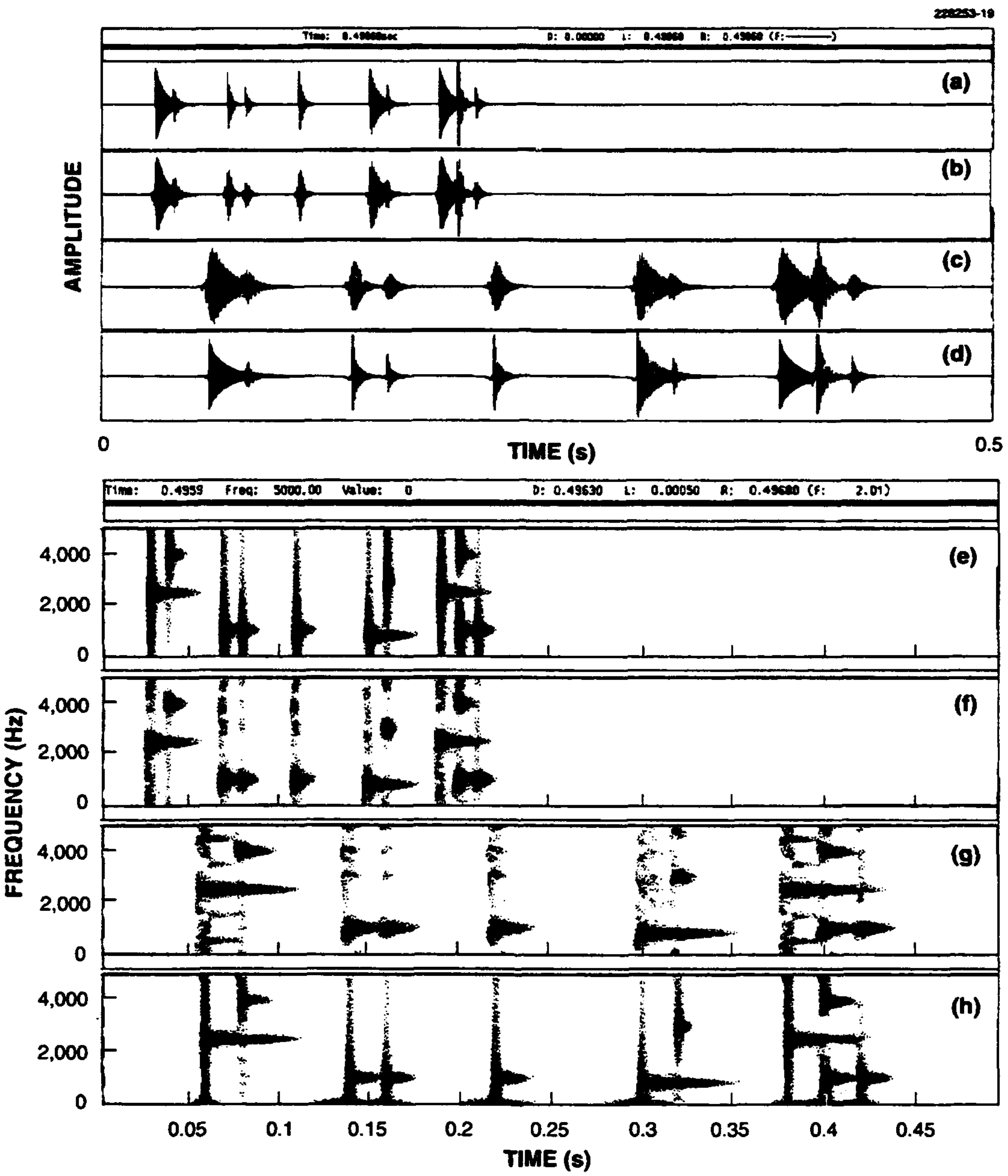

Figure A-3. Comparison of subband and sine-wave-based time-scaled expansion (×2): (a) original; (b) sine-wave reconstruction—3-ms window, 1-ms frame; (c) sine-wave expansion—3ms window, 1-ms frame; and (d) subband expansion. [(e)-(h) are spectrograms of (a)-(d).]

A.3 Background Preservation by Waveform Subtraction

When the object is tonal, an alternate approach taken by Serra [18] is to use sine-wave analysis/synthesis to estimate and subtract the object $x(n)$ from $s(n)$. The residual signal resulting from the subtraction is a background estimate that can be time-scaled according to Equation (24) while the tonal object is time-scaled using sine-wave analysis/synthesis. When the SNR is low, this approach is taxed by the accuracy required for tone subtraction. Even when the SNR is high, problems may arise in tones with sharp attacks. Nevertheless, an advantage of using this scheme in contrast to the subband approach of Section 3 is that the background is reconstructed uniformly over the fullband.

As a stepping stone to automating the procedure for tone extraction, an interactive (Matlab-based) system has been developed that allows a user to manually extract tones from $s(n)$. The system first displays the sine-wave tracks and the user then selects (i.e., clicks using a mouse) tracks associated with tones. These components are saved and subtracted from the signal $s(n)$ and modification is then applied (See Figure A-4). Figure A-5 illustrates an example of tone selection from a singing voice. For tonal object components this approach has been quite effective for time-scale modification.

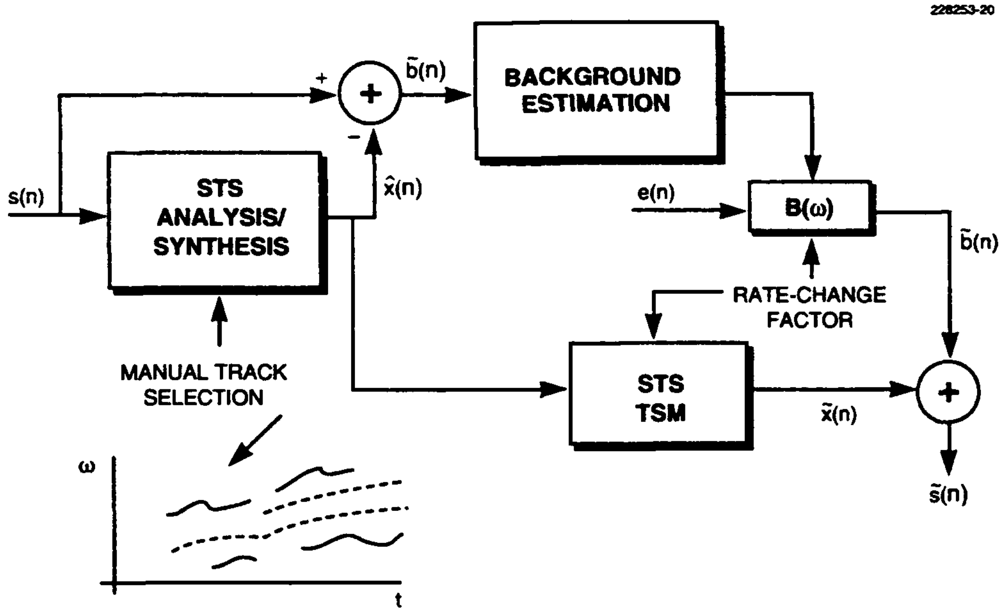

Figure A-4. Waveform subtraction algorithm for time-scale modification.

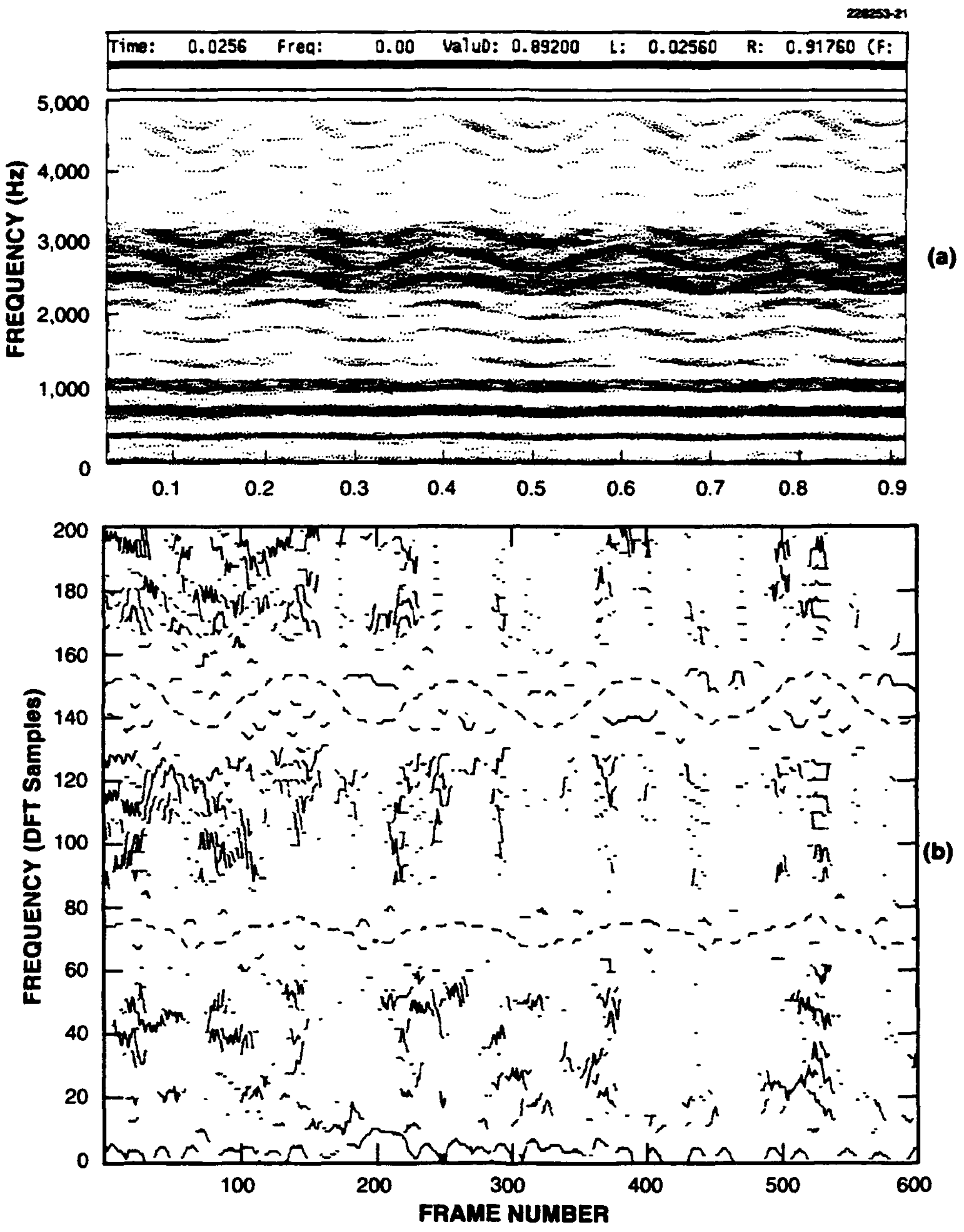

Figure A-5. Interactive tone selection using sine-wave tracks: (a) spectrogram of singing voice with vibrato and (b) selected tracks (dashed lines).

A.4 Background Preservation by Harmonic Sine-Wave Dithering

The proper way to perform phase dithering accounts for temporal correlation; the phase of a single sine wave should be dithered so that the resulting spectrum is narrowband and is concentrated near the carrier frequency of each sine wave. Sine-wave amplitudes might also be dithered; however, amplitude dithering was found to degrade the background quality and modify its temporal structure.

One method of phase dithering creates a dither function by passing a white-noise process through a low-pass filter with a bandwidth that is roughly the average distance between sine-wave frequencies. The problem with this approach is that sine-wave frequencies are nonuniformly spaced and change with time so that artificially high and low spectral concentrations can be built up. An alternative is to constrain sine waves to be harmonically related in regions that are considered background, making suitable the use of a fixed bandwidth.[9]

A.4.1 Phase-Dithering Algorithm

In this alternative sine-wave representation, sine-wave parameters are derived not by peak-picking, but by sampling the STFT at harmonically related frequencies

$$x_k(mL) = a_k(mL)\exp[j\Theta_k(mL)] = S(\omega_o k, mL) \quad . \tag{A.4}$$

The signal could then be synthesized and modified using a sine-wave synthesis of the form

$$s(n) = \sum_k A_k(n)\cos[\Theta_k(n)] \quad . \tag{A.5}$$

However, the resulting signal suffers from tonality. To avoid tonality, a dithering function is derived as

$$\epsilon_k(n) = \tan^{-1}[y_i(n)/y_r(n)] \quad , \tag{A.6a}$$

where

[9] Dithering harmonic sine waves is essentially the mechanism of synthesizing unvoiced speech in sine-wave speech coding [10]; however in the coder an uncorrelated random phase dither is applied only once on each synthesis frame, rather than sample by sample. Another approach related to sine-wave dithering, and also developed in the speech coding context, is the work of Marques and Almeida [24] who used narrowband basis functions in the representation of unvoiced speech sounds.

$$y_r(n) = h_{lpf}(n) * e_r(n) \tag{A.6b}$$

$$y_i(n) = h_{lpf}(n) * e_i(n) \quad , \tag{A.6c}$$

where $h_{lpf}(n)$ is a low-pass filter with bandwidth equal to the fundamental frequency ω_o, and $e_r(n)$ and $e_i(n)$ are white-noise inputs. The resulting phase-dithered signal is given by

$$s(n) = \sum_k A_k(n)\cos[\Theta_k(n) + \epsilon_k(n)] \quad . \tag{A.7}$$

Although the original .ine-wave parameters are computed on a frame basis, the phase dithering occurs sample by sample, which is necessary to obtain the proper temporal correlation in the resulting random process.

To determine when to phase dither, i.e., when the background is present, a scheme similar to the detection algorithm of Section 4.2 is used. The harmonic frequencies are selected to fall at the center frequencies of the bandpass filters used in white-noise subband shaping. However, for each frame m, the instantaneous energy required in the detection statistic $D_k(m)$, rather than being measured from subband filter outputs, is taken as the sum of the squared sine-wave amplitudes weighted by the respective subband filter. The phase dithering procedure is written as

$$\epsilon_k(n) = \tan^{-1}[y_i(n)/y_r(n)] \text{ if } D_k(m) \leq T \tag{A.8a}$$

and

$$\epsilon_k(n) = 0 \text{ if } D_k(m) > T \tag{A.8b}$$

Figure A-6 gives an overview of the modification algorithm.

The advantage of the algorithm is that it works solely in the sine-wave domain and, as shown below, it appears to preserve the temporal structure of the background. The disadvantage is that because only one function controls the dithering of all background sine waves, some spectral distortion results in the form of tonality. Because the phase dithering function is copied to each harmonic, the tonality is perceived as the fundamental frequency, or pitch, of the harmonic

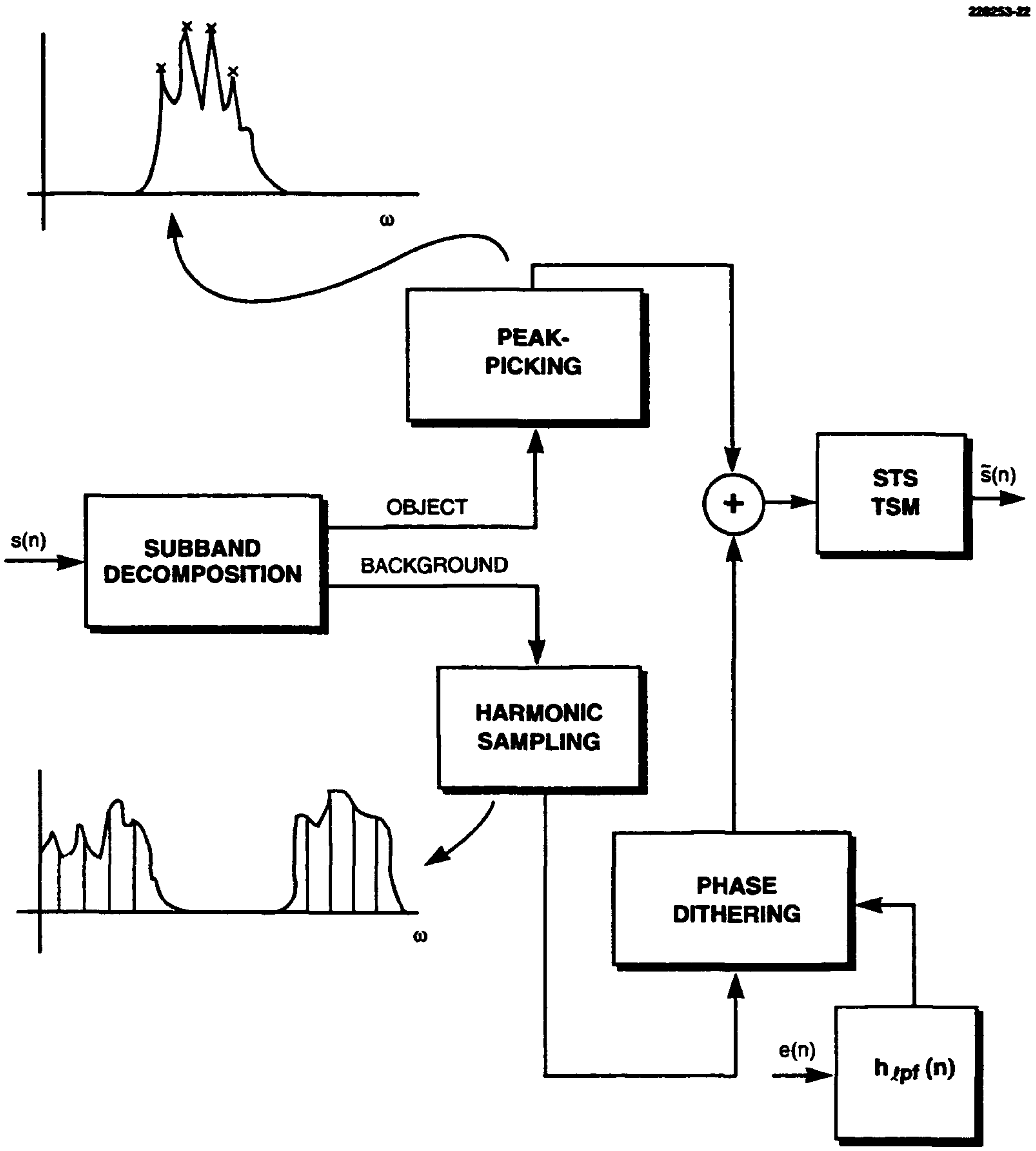

Figure A-6. Harmonic sine-wave dithering algorithm for time-scale modification.

structure. Nevertheless, this tonal distortion is far less objectionable than that from the baseline sine-wave modification of Equation (A.2).[10]

A.4.2 Examples

Figure A-7 illustrates the harmonic dithering scheme through an example. Twenty one harmonic sine waves were used as this is the number of bands in the subband system (Section 4). The background is colored noise generated by passing white Gaussian noise through a linear filter with two resonances at 2,500 and 4,000 Hz. The object component is taken from Figure 13. The signal-to-noise level was selected to make visible the temporal structure of the noise, hiding the object's tones and clicks. The figure shows that the harmonic dithering approach is capable of preserving the temporal envelope of the noise (while the subband filtering loses the temporal structure). An attempt is being made to obtain a better understanding of this property that stems likely from the use of the original sine-wave carrier phase $\Theta_k(n)$ and amplitudes $A_k(n)$. Spectrograms show that the spectral structure of the object and background are also largely preserved (at least visually) in the modification.

A.5 Discussion

A number of approaches are being considered for improving the object modification. One approach applies phase synchronization across sine waves. This concept of synchronizing sine waves was first used in the speech context [9,25,26] and in fact was the motivation for the subband phase synchronization of Section 4. Another approach imposes a fullband temporal envelope on the modified waveform, consisting of both object and background, by replacing the envelope of the modified waveform with a time-scaled version of the original temporal envelope (which can be considered as one iteration of the method described in Appendix B). Preliminary experiments indicate the technique aurally sharpens the object reconstruction.

[10] If a separate dithering function were to be used in each band, derived from distinct low-pass filtered white-noise processes, then the approach is similar (and perhaps under certain conditions can be shown to be identical) to the subband approach to background synthesis of Section 4.

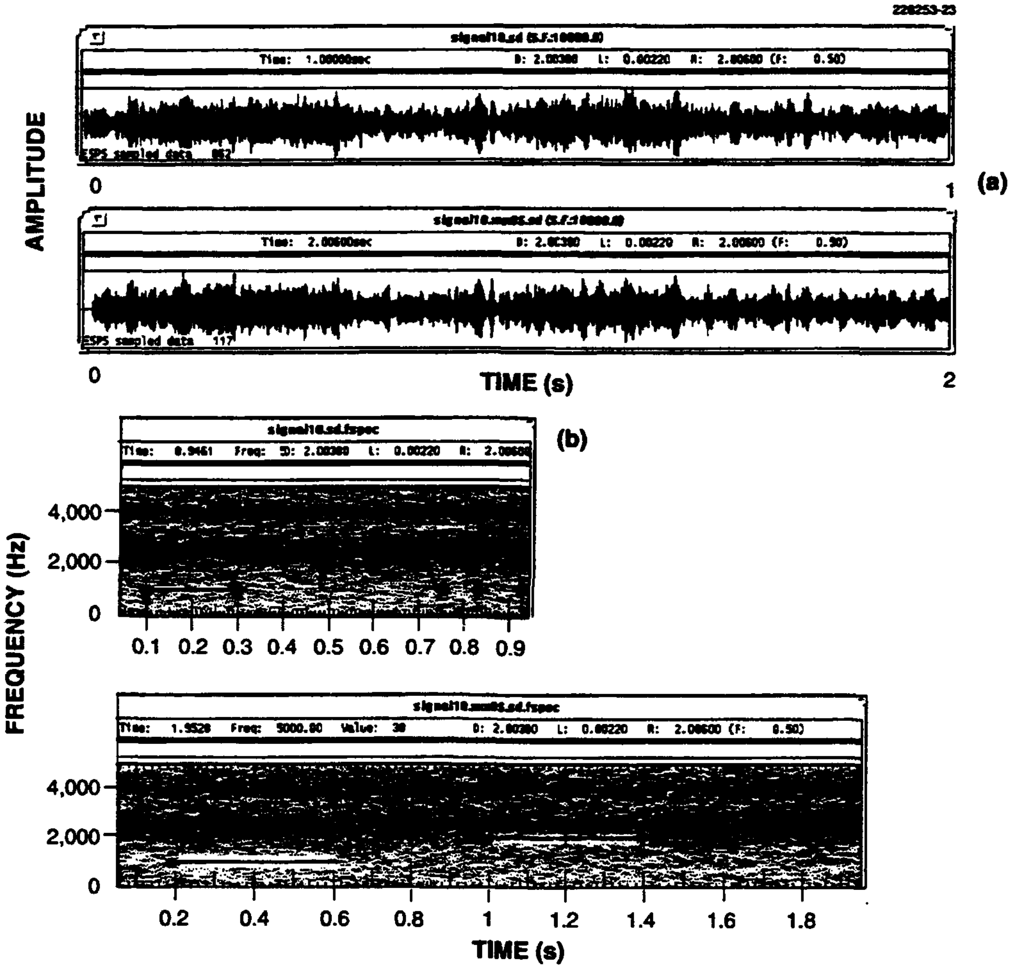

Figure A-7. Time-scale modification using harmonic sine-wave dithering: (a) waveforms and (b) spectrograms.

APPENDIX B
Iterative Algorithm for Time-Scale Modification

This appendix describes an iterative algorithm for time-scale modification of complex acoustic signals [14]. It is first shown that a signal can be iteratively reconstructed from its spectral and temporal envelope. This reconstruction technique is then extended to time-scale expand a signal using its spectral envelope and a time-scaled version of its temporal envelope.

B.1 Reconstruction from Temporal and Spectral Envelopes

Under certain conditions a signal can be represented by its partial specification in either the time domain, or the frequency domain, or both. For the purpose of time-scale modification, a signal is specified by its spectral and temporal envelopes. The definition of *spectral envelope* is the magnitude of the Fourier transform. One definition of the *temporal envelope* reflects the distinctive events of the complex acoustic signal, e.g., the start and stop time of a click or the modulation pattern of two beating sine waves. In the context of bandpass signals, temporal envelope is defined as the magnitude of an analytic signal representation (computed using the Hilbert Transform [11]); yet another definition uses measurements of attack and decay dynamics [12]. The advantage of the former approach is that it allows for a formal framework in which a signal can be represented by its temporal and spectral envelope. The latter approach has the advantage that the temporal envelope can be defined to reflect specific events.

Consider defining the temporal envelope to be the magnitude of the analytic signal. A discrete-time signal $x(n)$ is given in analytic form by

$$s(n) = x(n) + j\hat{x}(n) \quad , \tag{B.1a}$$

where $\hat{x}(n)$ is the Hilbert transform of $x(n)$, which in polar form is written as

$$s(n) = a(n)\exp[j\phi(n)] \quad , \tag{B.1b}$$

where $a(n)$ is the temporal envelope of the sequence. The Fourier transform of (B.1a) is given by

$$\begin{aligned} S(\omega) &= X(\omega) \qquad 0 \le \omega \le \pi \\ &= 0 \qquad\quad\; -\pi \le \omega < 0 \end{aligned} \tag{B.2a}$$

and is written in polar form as

$$S(\omega) = A(\omega)\exp[j\theta(\omega)] \quad , \tag{B.2b}$$

where $A(\omega)$ is the spectral envelope of the sequence.

For a large class of sequences $x(n)$, reconstruction can be performed iteratively from the temporal envelope $a(n)$ and the spectral envelope $A(\omega)$. In one iterative algorithm, the desired temporal envelope is imposed in the time domain via the Hilbert transform, and the specified spectral envelope is imposed in the frequency domain via the Fourier transform.[11] Specifically, on the kth iteration of the algorithm, the analytic representation of the signal estimate $x_k(n)$ is obtained in polar form as $s_k(n) = a_k(n) \exp[j\phi_k(n)]$. The temporal envelope of this estimate is replaced by the desired function $a(n)$

$$\hat{s}_k(n) = a(n)\exp[j\phi_k(n)] \quad . \tag{B.3}$$

The spectral envelope of the Fourier transform of $\hat{s}_k(n)$, $\hat{A}_k(\omega)\exp[j\hat{\theta}_k(\omega)]$, is then replaced by the specified spectral envelope

$$\tilde{S}_k(\omega) = A(\omega)\exp[j\hat{\theta}_k(\omega)] \quad . \tag{B.4}$$

The successive iterate $x_{k+1}(n)$ is formed by taking the real part of the inverse Fourier trans. orm of (B.4). In practice, the continuous Fourier transform is replaced by a discrete Fourier transform (implemented with an FFT), which is sufficiently long to avoid aliasing.

Empirically, it was found that the mean-squared error in the temporal envelope, $\sum_n[a_k(n) - a(n)]^2$, is nonincreasing as a function of iteration. The algorithm was demonstrated by reconstructing a variety of signals (consisting of damped sine waves) from initial estimates $s_0(n) = a_0(n)\exp[j\phi_0(n)]$, where the desired temporal envelope was used as the initial temporal envelope $a_0(n)$ and the initial phases $\phi_0(n)$ were assigned random values. Convergence was achieved within a few hundred iterations. A: example of reconstructing a sequence from its temporal and spectral envelopes is illustrated in Figure B-1. The synthetic signal, which was selected to resemble the response from a closing stapler, is the sum of two damped sine waves of different frequencies, displaced in time by 10 ms. The mean-squared error in the temporal envelope as a function of iteration is shown in Figure B-1(c).

[11] A similar iterative algorithm was recently (independently) proposed for peak-to-root-mean-square reduction where the desired temporal envelope is constant over signal duration [27]. However, unlike the general case, for constant temporal envelope there exists an approximate closed-form solution to the reconstruction [28].

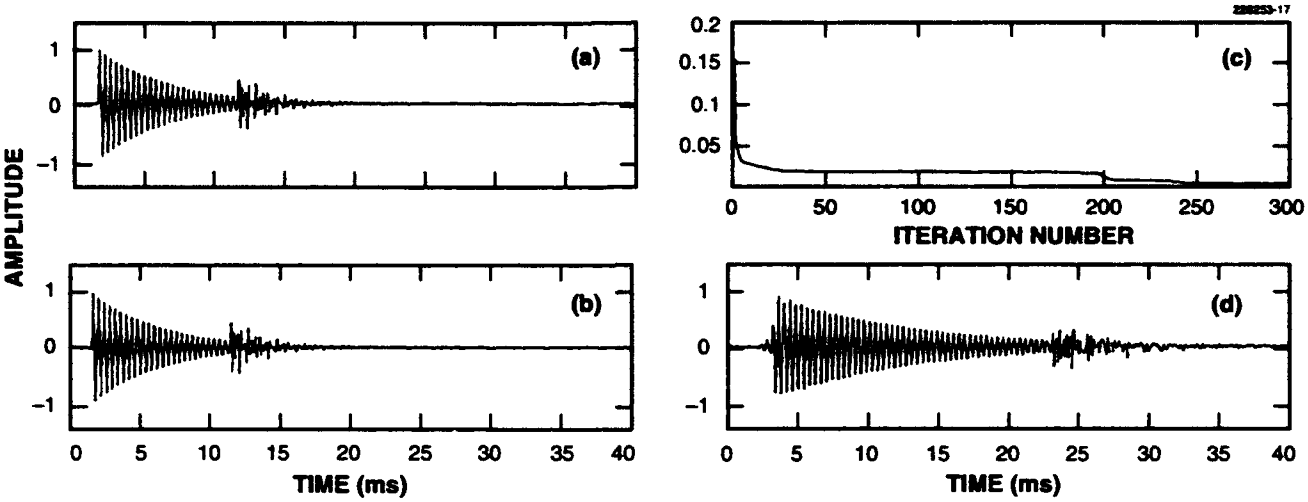

Figure B-1. Iterative reconstruction and modification from temporal and spectral envelopes: (a) original, (b) reconstruction, (c) temporal envelope error, and (d) time-scaled expansion.

B.2 Time-Scale Modification Using the Fullband Envelope

The essence of the approach to time-scale modification is to select a Fourier-transform phase which, when combined with a specified spectral envelope, results in a sequence with a time-scaled version of the original temporal envelope. However, a close match of both the specified spectral envelope and modified temporal envelope may not be "consistent" with the relationship between a sequence and its Fourier transform. Consider, for example, an exponentially damped sine wave. Expansion of the temporal envelope by slowing its rate of decay will narrow the signal's resonant bandwidth; but arbitrarily slowing the decay rate and maintaining the original resonant bandwidth may violate constraints on the signal's time and frequency concentrations [13]. In general, there may not necessarily exist a sequence jointly satisfying temporal and spectral envelope constraints.[12] The signal modification problem must therefore be formulated as finding a Fourier-transform phase that results in a sequence with a temporal envelope $\hat{\alpha}(n)$ that is close, in some sense, to a desired time-scaled envelope $\alpha(n)$.

One criterion of closeness is the mean-squared error between the modified temporal envelope $\alpha(n)$ and its estimate $\hat{\alpha}(n)$. The Fourier transform phase of the time-scaled sequence is chosen to minimize the mean-squared error

[12] However, in the special case where the signal's spectrum is constant a desired temporal envelope can be approximately matched [28].

$$E = \sum_{n} [\hat{\alpha}(n) - \alpha(n)]^2 \qquad \text{(B.5)}$$

This minimization is subject to the constraint that the spectral envelope of the modified sequence equals either $A(\omega)$ or a version of $A(\omega)$, which is more consistent with $\alpha(n)$ but which does not compromise the "spectral character" of the original sequence. The iterative reconstruction algorithm of the previous section represents a (possibly suboptimal) approach to solving this nonlinear optimization problem.

An example of the above modification algorithm is illustrated in Figure B-1(d) where the sequence from Figure B-1(a) has been time-scale expanded by a factor of 2 using 300 iterations. The desired temporal envelope $\alpha(n)$ was obtained by upsampling the original temporal envelope $a(n)$. The spectral envelope was left intact except for a mild spectral (resonant) sharpening of $A(\omega)$ to improve consistency with $\alpha(n)$. The mean-squared error between the desired and estimated temporal envelopes does not approach zero in this case. Nevertheless, the two closely spaced and overlapping components of the signal, which were barely audible in the original, are perceptually distinct in the modified signal.

B.3 Discussion

Although the capability of the iterative algorithm to recover a desired signal from temporal and spectral envelopes has been demonstrated empirically for some signals, it remains to prove signal uniqueness from such information. In addition, a number of potential problems need to be addressed; one is the possibility of obtaining local minima through the iteration. Another problem is the computational complexity (sometimes requiring hundreds of iterations), and thus methods to quicken convergence should be sought. Finally, but most importantly, the difficulty in obtaining a meaningful fullband envelope through the analytic signal representation or other approaches, for a general signal class, has yet to be resolved.

REFERENCES

1. T.E. Hanna, "Contributions of Envelope Information to Classification of Brief Sounds," NSMRL Report No. 1165, Groton, Conn.: Naval Submarine Medical Research Laboratory, (1990).

2. T.E. Hanna and Y.R. Masakowski, "Narrowband and Broadband Envelope Cues for Aural Classification," NSMRL Report No. 1171, Groton, Conn.: Naval Submarine Medical Research Laboratory (1991).

3. J.H. Howard, J.C. Solinsky, and M.H. Miller, "Classification of Acoustic Transients by Human Listeners," presented at the ONR Bioacoustic Signal Classification Workshop, Wilmington, N.C. (April 14–17, 1992).

4. D.J. Van Tasell, S.D. Soli, V.M. Kirby, and G.P. Widin, "Speech Waveform Envelope Cues for Consonant Recognition," *JASA*, **82** (1987).

5. M. Dolson, "The phase vocoder: A tutorial," *Computer Music J.*, **10** (1986).

6. T.F. Quatieri and R.J. McAulay, "Speech transformations based on a sinusoidal representation," *IEEE Trans. Acoust. Speech Signal Process.*, **34** (1986).

7. R.J. McAulay and T.F. Quatieri "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust. Speech Signal Process.*, **40** (1986).

8. H. Nawab and T.F. Quatieri, "Short-Time Fourier Transform," in *Advanced Topics in Signal Processing*, J.S. Lim and A.V. Oppenheim, eds., Englewood Cliffs, N.J.: Prentice Hall (1988).

9. T.F. Quatieri and R.J. McAulay, "Shape-invariant time-scale and pitch modification of speech," *IEEE Trans. Acoust. Speech Signal Process.*, **40** (1992).

10. R.J. McAulay and T.F. Quatieri "Low Rate Speech Coding Based on the Sinusoidal Speech Model," in *Recent Progress in Speech Signal Processing*, S. Furui and M.M. Sondhi, eds., New York: Marcel Dekker (1992).

11. A.V. Oppenheim and R. Schafer, *Digital Signal Processing*, Englewood Cliffs, N.J.: Prentice Hall (1975).

12. B.A. Blesser, "Audio dynamic range compression for minimum perceived distortion," *IEEE Trans. Audio Acoust.*, **AU-17** (1969).

13. A. Papoulis, *The Fourier Integral and Its Applications*, New York, N.Y.: McGraw Hill (1962).

14. T.F. Quatieri, R.B. Dunn, and T.E. Hanna, "Time-Scale Modification of Complex Acoustic Signals," Minneapolis, Minn.: *Proc. IEEE Conf. on Acoust. Speech Signal Process.* (April 1993).

REFERENCES
(Continued)

15. T.F. Quatieri, R.B. Dunn, and T.E. Hanna, "Time-Scale Modification with Temporal Envelope Invariance," New Paltz, N.Y.: *Proc. of IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics* (October 1993).

16. S. Mallat and W.L. Hwang, "Singularity detection and processing with wavelets," *IEEE Trans. Inf. Theory*, **38** (1992).

17. S. Roucos and A.M. Wilgus, "High Quality Time-Scale Modification for Speech," Tokyo: *Proc. IEEE Conf. on Acoust. Speech Signal Process.* (April 1986).

18. X. Serra, "A System for Sound Analysis/Transformation/Synthesis Based on a Deterministic Plus Stochastic Decomposition," Ph.D. Thesis, CCRMA, Department of Music, Stanford University (1989).

19. T.F. Quatieri, "Object Detection by Two-Dimensional Linear Prediction," Boston, Mass.: *Proc. IEEE Conf. on Acoust. Speech Signal Process.* (April 1983).

20. T.F. Quatieri, "Object Detection by Two-Dimensional Linear Prediction," MIT Lincoln Laboratory Technical Rep. 632 (January 1983), DTIC AD-A126340/9.

21. J. Laroche, Y. Stylianou, and E. Moulines, "HNS: Speech Modification Based on a Harmonic + Noise Model," Minneapolis, Minn.: *Proc. IEEE Conf. on Acoust. Speech Signal Process.* (April 1993).

22. A.S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*, Cambridge, Mass.: The MIT Press (1990).

23. D.M. Rubin, "Use of forecasting signatures to help distinguish periodicity, randomness, and chaos in ripples and other spatial patterns," *Chaos*, **2** (1992).

24. J.S. Marques and L.B. Almeida, "Sinusoidal modeling of speech: Representation of unvoiced sounds with narrowband basis functions,"*Proc. EUSIPCO* (1988).

25. R.J. McAulay and T.F. Quatieri "Phase Modeling and its Application to Sinusoidal Transform Coding," Tokyo: *Proc. IEEE Conf. on Acoust. Speech Signal Process.* (1986).

26. T.F. Quatieri and R.J. McAulay, "Phase Coherence in Speech Reconstruction for Enhancement and Coding Applications," Glasgow: *Proc. IEEE Conf. on Acoustics, Speech, and Signal Processing* (May 1989).

27. J. Cartinhour, "An iterative time-frequency domain algorithm for reduction of peak-to-RMS ratio," *Digital Signal Processing*, 236–241 (1992).

28. T.F. Quatieri, J.T. Lynch, M.L. Malpass, R.J. McAulay, and C.W. Weinstein, "Peak-to-RMS reduction based on a sinusoidal model," *IEEE Trans. Acoust. Speech Signal Process.*, **39** (1991).

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (*Leave blank*)

2. REPORT DATE
4 February 1994

3. REPORT TYPE AND DATES COVERED
Technical Report

4. TITLE AND SUBTITLE
Time-Scale Modification of Complex Acoustic Signals in Noise

5. FUNDING NUMBERS
C — F19628-90-C-0002
PR — 411

6. AUTHOR(S)
Thomas F. Quatieri, Robert B. Dunn, Robert J. McAulay, and Thomas E. Hanna

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Lincoln Laboratory, MIT
P.O. Box 73
Lexington, MA 02173-9108

8. PERFORMING ORGANIZATION REPORT NUMBER
TR-990

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
NSMRL
Box 900, Subbase NLON
Groton, CT 06349

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
ESC-TR-93-263

11. SUPPLEMENTARY NOTES
None

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT (*Maximum 200 words*)

A new approach is introduced for time-scale modification of short-duration complex acoustic signals to improve their audibility. The method preserves an approximate time-scaled temporal envelope of a signal, thus capitalizing on the perceptual importance of the signal's temporal structure, while also maintaining the character of a noise background. The basis for the approach is a subband signal representation, derived from a filter bank analysis/synthesis, the channel phases of which are controlled to shape the temporal envelope of the time-scaled signal. Channel amplitudes and filter bank inputs are selected to shape the spectrum and correlation of the time-scaled background. The phase, amplitude, and input control are derived from locations of events that occur within filter bank outputs. A frame-based generalization of the method imposes phase consistency and background noise continuity across consecutive synthesis frames. The approach and its derivatives are applied to synthetic and actual complex acoustic signals consisting of closely spaced sequential time components.

14. SUBJECT TERMS
time-scale modification
signal enhancement
temporal envelope
subband representation
slow-motion audio replay
noise background preservation
filter bank analysis/synthesis
complex acoustic signal
improved audibility
sine-wave analysis/synthesis

15. NUMBER OF PAGES
70

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT
Unclassified

18. SECURITY CLASSIFICATION OF THIS PAGE
Unclassified

19. SECURITY CLASSIFICATION OF ABSTRACT
Unclassified

20. LIMITATION OF ABSTRACT
Same as Report