Improved Phase Vocoder Time-Scale Modification of Audio

Jean Laroche, *Member, IEEE,* and Mark Dolson, *Member, IEEE*

***Abstract*—The phase vocoder is a well-established tool for time scaling and pitch shifting speech and audio signals via modification of their short-time Fourier transforms (STFT's). In contrast to time-domain time-scaling and pitch-shifting techniques, the phase vocoder is generally considered to yield high quality results, especially for large modification factors and/or polyphonic signals. However, the phase vocoder is also known for introducing a characteristic perceptual artifact, often described as "phasiness," "reverberation," or "loss of presence." This paper examines the problem of phasiness in the context of time-scale modification and provides new insights into its causes. Two extensions to the standard phase vocoder algorithm are introduced, and the resulting sound quality is shown to be significantly improved. Moreover, the modified phase vocoder is shown to provide a factor-of-two *decrease* in computational cost.**

***Index Terms*—Phase coherence, phase vocoder, pitch shifting, short time Fourier transform, time scaling.**

I. Introduction

TIME-SCALE and pitch-scale modification of signals has long been a subject of interest in the audio and speech processing community. In recent years, though, there has been a dramatic increase in commercial application of these techniques. Time-scaling and/or pitch-shifting algorithms are now being used in a widening array of devices such as telephone answering systems, musical effect processors, professional CD players, hard-disk recorders, PC-based sound editors, and so on. As the computational resources of these devices increase, so too do expectations for their audio fidelity.

Most commercial implementations of time scaling (or pitch shifting)[1] use time-domain-based techniques which rely upon some form of synchronized overlap-add of signal excerpts. These methods are attractive for their relatively low computational cost and because they yield good results in some special cases of interest (e.g., modification factors close to one or monophonic sounds). However, these techniques tend to perform poorly when applied to complex, polyphonic, or nonpitched signals, or when large modification factors must be used (e.g., factors greater than $\pm 20\%$ to $\pm 30\%$). In these cases, typical artifacts include warbling (a type of periodic frequency modulation observed in processed polyphonic signals), transient doubling or skipping (especially troublesome for percussive signals), and tempo modulation. A full discussion of time-domain time-scaling techniques and their shortcomings can be found in [7] or [5].

In contrast, frequency-domain-based time-scaling techniques such as the phase vocoder [1] employ a fixed overlap-add approach; synchronization between overlapping frames is obtained by modifying phases in the signal's short-time Fourier transform. Phase vocoder time-scaling is not limited to near-unity modification factors nor to monophonic signals, and it is free of many of the typical time-domain artifacts.[2] This makes it a potentially attractive approach. However, the computational cost of the phase vocoder is much higher than that of time-domain techniques *and* the algorithm introduces distinctive artifacts of its own. Ultimately, it is these artifacts that pose the major barrier to more widespread use of the phase vocoder.

The two most prominent phase vocoder time-scaling artifacts are "transient smearing" and "phasiness." Transient smearing occurs even with modification factors that are close to one, and is heard as a slight loss of percussiveness in the signal; piano attacks, for example, may be perceived as having less "bite." Phasiness (or reverberation or "loss of presence") also occurs even with near-unity modification factors, and is heard as a characteristic coloration of the signal; in particular, time-expanded speech often sounds as if the speaker is much further from the microphone than in the original recording.

In general, neither time-domain nor frequency-domain time-scaling artifacts have received much attention in the technical literature, probably because assessments of fidelity have varied according to local standards (and over time as well). The problem of transient smearing in subband-based time-scale modifications is addressed in [11] and an improved technique based on phase-locking at transient times is proposed. The phenomenon of phasiness has been noted by several authors [10], [14], and the root of the problem is known to lie in the modification of phases in the short-time Fourier transform (STFT). To date, however, no thorough explanation for this phenomenon has yet been given, and proposed solutions have proven to be either cumbersome or only marginally effective.

This paper proposes an explanation for the presence of phasiness in time-scaled signals, and offers new phase calculation techniques that are shown to significantly reduce the problem. In addition, these new techniques make it possible to reduce the computational cost of the phase vocoder by more than a factor of two.

Manuscript received June 9, 1997; revised May 14, 1998. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Juergen H. Herre.

The authors are with Joint E-mu/Creative Technology Center, Scotts Valley, CA 95067 USA (e-mail: jeanl@emu.com markd@emu.com).

Publisher Item Identifier S 1063-6676(99)02745-5.

[1] Since pitch-scale modification can be performed by combining time scaling and sample-rate conversion, we focus in this paper exclusively on time scaling.

[2] For example, tempo modulation is virtually nonexistent, and warbling can be eliminated by an appropriate choice of parameters.

The remainder of this paper is divided into two sections. In the first part, the emphasis is on understanding the problem: the standard phase-vocoder technique for time scaling is described, and a detailed investigation of potential phase errors is presented. In the second part, the focus is on solutions: two previously-proposed solutions are briefly reviewed, and two new phase-modification techniques are introduced and evaluated.

II. The Basic Phase Vocoder Time-Scaling Algorithm

The essence of time scaling is the modification of a signal's temporal evolution while its local spectral characteristics are kept unchanged. Phase-vocoder-based time-scaling techniques accomplish this via an explicit sequence of analysis, modification, and resynthesis.

A. Phase Vocoder Analysis/Synthesis

During the analysis stage, analysis time-instants t_a^u for successive values of integer u are set along the original signal, possibly uniformly: $t_a^u = uR_a$ where R_a is the so-called analysis hop factor. At each of these analysis time-instants, a Fourier transform is calculated over a windowed portion of the original signal, centered around t_a^u. The result is the *nonheterodyned* STFT representation of the signal, denoted $X(t_a^u, \Omega_k)$:

$$X(t_a^u, \Omega_k) = \sum_{n=-\infty}^{\infty} h(n)x(t_a^u + n)e^{-j\Omega_k n} \tag{1}$$

where x is the original signal, $h(n)$ is the analysis window, $\Omega_k = \frac{2\pi k}{N}$ is the center frequency of the kth vocoder "channel," and N is the size of the discrete Fourier transform. In practice, $h(n)$ has a limited time span (typically N samples) and the sum above has a finite number of terms. $X(t_a^u, \Omega_k)$ is both a function of time (via variable u) and frequency (via Ω_k).

The resynthesis stage involves setting synthesis time-instants t_s^u, usually uniformly, so that $t_s^u = R_s u$, where R_s is the synthesis hop factor. At each of these synthesis time-instants, a short-time signal $y_u(n)$ is obtained by inverse-Fourier-transforming the synthesis STFT $Y(t_s^u, \Omega_k)$. Each short-time signal is then multiplied by an optional synthesis window $w(n)$, and the windowed short-time signals are all summed together, yielding the output signal $y(n)$:

$$\begin{aligned} y(n) &= \sum_{u=-\infty}^{\infty} w(n - t_s^u)y_u(n - t_s^u) \quad \text{with} \\ y_u(n) &= \frac{1}{N}\sum_{k=0}^{N-1} Y(t_s^u, \Omega_k)e^{j\Omega_k n}. \end{aligned} \tag{2}$$

In the absence of modifications (i.e., $R_a = R_s$ and $Y(t_s^u, \Omega_k) = X(t_a^u, \Omega_k)$), this output signal is identical to the original signal x, under mild conditions on the analysis and synthesis windows [4]. In general, however, a modified $Y(t_s^u, \Omega_k)$ is not the STFT of *any* actual signal. In particular, the output signal $y(n)$ obtained via the above reconstruction formula does *not* necessarily have $Y(t_s^u, \Omega_k)$ as its STFT. The sequence of STFT frames for a given signal must satisfy strong consistency conditions because the Fourier transforms correspond to *overlapping* short-time signals. The formula above merely yields a signal whose STFT is close to $Y(t_s^u, \Omega_k)$ in a sense that depends on the choice of the synthesis window $w(n)$. Further elaboration of this point can be found in [4], [7].

B. Time-Scale Modifications

In phase-vocoder-based time scaling, the STFT is modified in two ways: 1) the analysis hop factor R_a is different from the synthesis hop factor R_s, and 2) the phase values of the synthesis STFT $Y(t_s^u, \Omega_k)$ are calculated explicitly according to a formula given below. These modifications are based on an underlying sinusoidal signal model, but no explicit parametric sinusoidal estimation is performed.

According to the underlying model, the input signal is the sum of a number $I(t)$ of sinusoids with time-varying amplitudes $A_i(t)$ and instantaneous frequencies $\omega_i(t)$

$$\begin{aligned} x(t) &= \sum_{i=1}^{I(t)} A_i(t)e^{j\phi_i(t)} \quad \text{with} \\ \phi_i(t) &= \phi_i(0) + \int_0^t \omega_i(\tau)\,d\tau \end{aligned} \tag{3}$$

in which $\phi_i(t)$ and $\omega_i(t)$ are called the instantaneous phase and frequency of the ith sinusoid.

Based on (3), for a constant modification factor α such that $t_s^u = \alpha t_a^u$, the ideal synthesis phase $\phi_s(t_s^u)$ of the time-scaled sinusoid i would be

$$\begin{aligned} \phi_s(t_s^u) &= \phi_s(0) + \int_0^{t_s^u} \omega_i(\tau/\alpha)\,d\tau \\ &= \phi_s(0) + \alpha\int_0^{t_a^u} \omega_i(\tau)\,d\tau \\ &= \phi_s(0) + \alpha[\phi_i(t_a^u) - \phi_i(0)] \end{aligned} \tag{4}$$

where $\phi_s(0)$ is an arbitrary initial synthesis phase.

Phase-vocoder-based time scaling modifies the STFT of the sinusoidal input signal components so as to produce the above time-scaled sinusoids. The time-evolution of the sine-wave amplitudes is modified simply by setting $|Y(t_s^u, \Omega_k)| = |X(t_a^u, \Omega_k)|$ where $t_s^u = R_s u$. However, modification of the sine-wave phases is more challenging.

To calculate the phase of $Y(t_s^u, \Omega_k)$, the standard phase-vocoder technique requires phase unwrapping, a process whereby the phase increment between two consecutive frames is used to estimate the instantaneous frequency of a nearby sinusoid in each channel. The instantaneous frequency $\hat{\omega}_k(t_a^u)$ is estimated by first calculating the *heterodyned* phase increment

$$\Delta\Phi_k^u = \angle X(t_a^u, \Omega_k) - \angle X(t_a^{u-1}, \Omega_k) - R_a\Omega_k$$

then taking its principal determination (between $\pm\pi$) denoted $\Delta_p\Phi_k^u$ and deriving the instantaneous frequency $\hat{\omega}_k(t_a^u)$ of the closest sinusoid using

$$\hat{\omega}_k(t_a^u) = \Omega_k + \frac{1}{R_a}\Delta_p\Phi_k^u. \tag{5}$$

This procedure is called *phase unwrapping*, because the actual (nonwrapped) value of the phase increment is calculated from its principal (wrapped) determination. The heterodyned phase increment $\Delta\Phi_k^u$ is simply the small phase shift resulting from $\omega_k(t_a^u)$ being close but not necessarily equal to Ω_k.

Once the instantaneous frequency at time t_a^u is estimated, the phase of the time-scaled STFT at time t_s^u is set according to the following *phase-propagation* formula

$$\angle Y\left(t_s^u, \Omega_k\right) = \angle Y\left(t_s^{u-1}, \Omega_k\right) + R_s \hat{\omega}_k\left(t_a^u\right). \tag{6}$$

Equation (6) guarantees what can be called "horizontal phase coherence": for a *constant-frequency* sinusoid, successive short-time signals will overlap coherently. Another way of saying this is that there is coherence *within* each frequency channel over time (i.e., along the horizontal dimension of a standard sonagram).

For constant-frequency sinusoids, the phase unwrapping (5) yields a good estimate of the instantaneous frequency if channel k is influenced by only one sinusoid, and if the analysis window's cutoff frequency ω_h is such that $R_a\omega_h < \pi$. In practice, for standard analysis windows (such as Hanning or Hamming windows), this constrains the analysis windows to overlap by at least 75%. The phase unwrapping equation involves the calculation of a four-quadrant arc tangent, and the phase propagation equation (6) requires the use of trigonometric functions in order to calculate the real and imaginary part of $Y(t_s^u, \Omega_k)$. An important point is that (6) does not indicate how the short-time Fourier phases should be set at the first synthesis time-instant t_s^0. As will be shown later, the choice of initial phases can significantly influence the quality of the output signal.

The technique outlined above is the standard frequency-domain time-scaling method based on the phase vocoder. Variants of this approach have also been proposed. Portnoff in [9] describes a technique applicable to speech signals, where the phases of the underlying sinusoidal components are decomposed into a system phase (resulting from the vocal tract filtering) and a source phase corresponding to glottal pulses. Interested readers can also refer to [10] for an alternative phase-updating method that does not involve phase unwrapping nor any trigonometric calculation, but instead uses an additional analysis Fourier transform.

C. Phase Problems in Phase-Vocoder Time-Scaling

1) Phase Coherence: Because phase propagation errors are at the heart of many of the sound quality issues in the phase vocoder, it is important to understand how sinusoidal phases are altered by vocoder-based time-scale modifications. The phase-vocoder time-scaling algorithm ensures phase consistency *within* each frequency channel over time, but it is also important to have phase consistency *across* the channels in a given synthesis frame. We call this latter requirement *vertical phase coherence*.

Both horizontal *and* vertical phase coherence must be preserved upon resynthesis for the STFT to be a "valid" one. If phase coherence is not preserved in the synthesis STFT $Y(t_s^u, \Omega_k)$, the synthesis equation (2) will yield a signal whose STFT is not close to $Y(t_s^u, \Omega_k)$. This new signal will likely exhibit beating in its individual harmonics, and this will be heard as phasiness or reverberation.

How can we recognize vertical phase coherence in a STFT? For a constant-amplitude, constant-frequency sinusoid, there is a simple phase relation between adjacent channels located around the sinusoidal frequency. If a sinusoid with a constant frequency ω_i falls in channel k, and if the analysis window $h(n)$ is symmetric around zero, it is easy to show that the channels around channel k which are influenced by this sinusoid (channels such that $|\Omega_k - \omega_i| < \omega_h$ where ω_h is the cutoff frequency of the analysis window) have an analysis phase equal to that of channel k. In practice, the analysis window $h(n)$ is more usually nonzero for $0 \le n < L$ and is symmetric around its middle point, but this changes things only slightly: If the size of the discrete Fourier transform is equal to the analysis window length L, then adjacent channels exhibit phase differences of $\pm\pi$.

For more complicated signals, unfortunately, no comparably simple phase relationship exists. For a sinusoid with a slowly varying frequency, the phases in channels around the instantaneous frequency are still nearly equal, but an analytical formula is difficult to develop. Consequently, there is no simple way to check for vertical phase coherence.

An *a posteriori* way to check the consistency of a STFT consists of reconstructing the synthesis signals $y(n)$, and checking that the STFT of $y(n)$ is indeed very close to $Y(t_s^u, \Omega_k)$ in both amplitude and phase. We propose a measure D_M of consistency derived from that proposed in [4]:

$$D_M = \frac{\sum_{u=P}^{U-P-1} \sum_{k=0}^{N-1} \left[\left|Z\left(t_s^u, \Omega_k\right)\right| - \left|Y\left(t_s^u, \Omega_k\right)\right|\right]^2}{\sum_{u=P}^{U-P-1} \sum_{k=0}^{N-1} \left|Y\left(t_s^u, \Omega_k\right)\right|^2} \tag{7}$$

where $Z(t_s^u, \Omega_k)$ is obtained by performing a STFT on the modified signal $y(n)$. U is the total number of short-time frames, and the summation excludes the P first and P last few frames to avoid taking into account errors due to missing overlapped segments in the resynthesis formula. The smaller D_M, the better the consistency. If total consistency is achieved, $Z(t_s^u, \Omega_k) = Y(t_s^u, \Omega_k)$ for all times t_s^u and all channels Ω_k and $D_M = 0$. Vertical and horizontal phase coherence will play an important role in the following sections, and we will use the measure above to estimate the degree of consistency of our algorithms.

2) Output Phase versus Input Phase: In this section, we seek to relate the phase of the modified STFT in channel k to the phase of the corresponding analysis STFT in the same channel. Assuming a constant modification factor $\alpha = \frac{R_s}{R_a}$, and given an initial synthesis Fourier transform phase $\phi_s(0, k)$, we can use the phase propagation equation (6) to express the phase of the output STFT at any given synthesis time-instant y_s^u. By iterating (6) for successive values of u, starting at $u = 0$, we obtain

$$\begin{aligned}\angle Y\left(t_s^u, \Omega_k\right) &= \angle Y\left(t_s^0, \Omega_k\right) + \sum_{i=1}^{u} R_s \hat{\omega}_k\left(t_a^i\right) \\ &= \phi_s(0, k) + \sum_{i=1}^{u} R_s \hat{\omega}_k\left(t_a^i\right)\end{aligned}$$

where $\hat{\omega}_k(t_a^i)$ is the estimated instantaneous frequency at time t_a^i in channel k. Now, using (5), we get

$$\angle Y(t_s^u, \Omega_k) = \phi_s(0,k) + \sum_{i=1}^{u} \left[R_s \Omega_k + \frac{R_s}{R_a} \Delta_p \Phi_k^i \right]$$

and using the definition of $\Delta_p \Phi_k^i$ we get

$$\angle Y(t_s^u, \Omega_k) = \phi_s(0,k) + \alpha \sum_{i=1}^{u} \left[\angle X(t_a^i, \Omega_k) - \angle X(t_a^{i-1}, \Omega_k) + 2m_k^i \pi \right]$$

where m_k^i is the unwrapping factor at the analysis time-instant t_a^i: $2m_k^i\pi = \Delta_p \Phi_k^i - \Delta \Phi_k^i$. This yields

$$\angle Y(t_s^u, \Omega_k) = \phi_s(0,k) + \alpha \left[\angle X(t_a^u, \Omega_k) - \angle X(0, \Omega_k) \right] + \alpha \sum_{i=1}^{u} 2m_k^i \pi. \quad (8)$$

Equation (8) gives the expression of the phase of the synthesis STFT at time t_s^u as a function of the synthesis initial phase $\phi_s(0,k)$, the phase of the analysis STFT at time t_a^u, the initial analysis phase, the modification factor α, and the series of phase-unwrapping integers m_k^i.

Several conclusions can be drawn from this equation.

- Equation (8) indicates that the synthesized phase depends on the analysis phases only at the current analysis time-instant and at the origin. This means that if an analysis phase is estimated incorrectly at any given time-instant, this error will not generate phase drift in *subsequent* frames, provided that the phase-unwrapping factor m_k^i remains correct.
- On the other hand, the series of phase-unwrapping factors m_k^i *do* have a cumulative effect: if an erroneous phase-unwrapping factor is calculated at a given frame, all subsequent frames will show a phase bias.
- Potential phase-unwrapping errors manifest themselves by multiples of $2\alpha\pi$ being added to the synthesis phase. If α is an integer, then phase-unwrapping errors are transparent since they always are multiples of 2π. As a result, integer-factor time-scaling operations can be performed *without phase unwrapping* by use of (8) where the factor $\sum_{i=1}^{u} 2m_k^i\pi$ is dropped. Skipping the phase-unwrapping stage significantly reduces the computation cost of such modifications.

Equation (8) also provides a solid analytical foundation for understanding the lack of vertical phase coherence in standard phase vocoder implementations. This issue is examined in detail in the next two sections.

3) Loss of Vertical Phase Coherence: According to (8), vertical phase coherence depends upon two factors: 1) initial phase values, and 2) accumulated phase-unwrapping errors. To see this, suppose at first that the modification factor α is a constant integer, so phase-unwrapping errors do not influence the modified signal. Now, consider a sinusoid whose instantaneous frequency varies across time so that it will migrate from channel to channel. Rearranging the terms in (8), we can express the synthesis phase of the peak channel at time t_s^u as

$$Y(t_s^u, \Omega_k) = \alpha \angle X(t_a^u, \Omega_k) + \theta_k \quad \text{with} \quad \theta_k = \phi_s(0,k) - \alpha \angle X(0, \Omega_k) \quad (9)$$

where the sum of unwrapping factors has been dropped, being a multiple of 2π. This expression differs from the ideal synthesis phase equation (4) in that θ_k is not necessarily a constant, but varies with the channel index k. The fact that θ_k may not be constant has two adverse and related consequences:

1) The synthesis phases in adjacent channels may be very different, if the values of θ_k vary significantly from channel to channel. As mentioned in II-B1, this should not be the case.
2) When the sinusoid's instantaneous frequency migrates from channel k_o at time t_a^u to channel $k_o + 1$ at time t_a^{u+1}, the synthesis phase undergoes a jump equal to $\theta_{k_o+1} - \theta_{k_o}$, which is also very undesirable.

As shown by (9) the values of θ_k for successive channels depend only on the analysis and the synthesis phases at time zero. If for example an area of the spectrum was dominated by noise at that time, the values of $\angle X(0, \Omega_k)$ and consequently of θ_k for the corresponding channels will be random, and are likely to exhibit large variations from channel to channel, unless we set the initial synthesis phase $\phi_s(0,k)$ such that

$$\theta_k = \phi_s(0,k) - \alpha \angle X(0, \Omega_k) = C \quad (10)$$

where C is a channel-independent constant. If the above equation is satisfied, then none of the two problems mentioned above occur, and the synthesis phase becomes identical to the ideal synthesis phase in (4).

Now, consider the more general case in which the modification factor α is not an integer, and again suppose that there is a sinusoidal component in the vicinity of channel k_o. Equation (8) indicates that, even if (10) is satisfied, phase coherence is guaranteed only if the sums of the unwrapping factors $\sum_{i=1}^{u} 2m_k^i\pi$ are equal (modulo 2π) in nearby channels. There is no danger of phase-unwrapping errors in channels near the sinusoid's instantaneous frequency so long as the analysis hop factor R_a is small enough. The deeper problem, though, is that no single sinusoidal component of an audio signal is likely to persist without interruption across the entire duration of the signal. Thus, there will inevitably be times during which channel k_o and its neighbors are influenced by unrelated sinusoids or even noise. It is during these times that phase-unwrapping differences will necessarily accumulate, making the terms $\sum_{i=1}^{u} 2m_k^i\pi$ different in adjacent channels; as a result, vertical phase coherence will quickly be lost forever.

Considering the universality of the above scenario, it seems truly amazing that the phase vocoder should work at all! For noninteger modification factors, vertical phase coherence is almost guaranteed to be lacking unless each sinusoidal component remains in the same phase vocoder channel for all time. Clearly, this assumption is violated by most signals of interest, including speech and music.

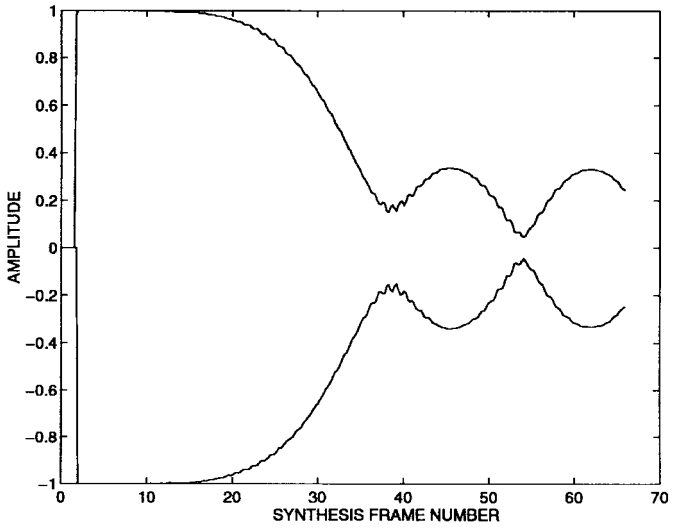

Fig. 1. Factor-two time-scaling of a constant amplitude chirp. Time-domain amplitude-envelope of the modified signal.

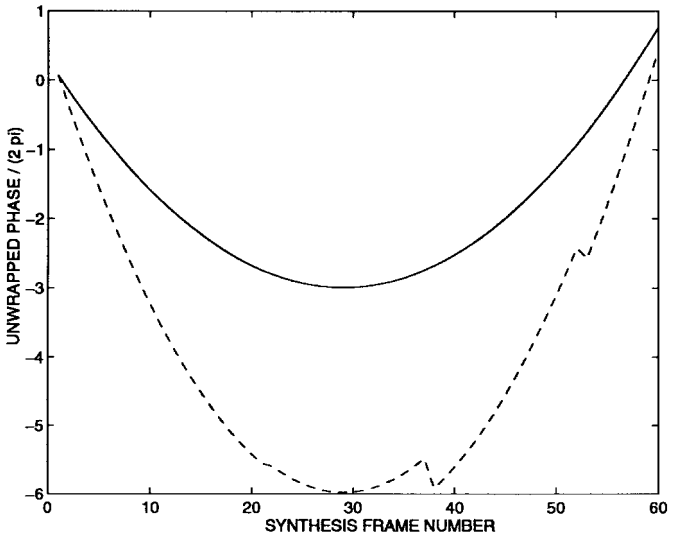

Fig. 2. Analysis phase (solid line) and synthesis phase (dotted line) in numbers of 2π, showing phase jumps at frames 22, 38, and 53.

4) Examples: In this section, we present a few time-scale modification examples that demonstrate the problems described above. We begin by noting that constant-frequency, constant-amplitude sinusoids pose no problem at all to the phase vocoder; the results are usually excellent, with consistency measures $D_M < -60$ dB, provided the initial synthesis phases are all set according to (10). Variable-amplitude sinusoids pose some problems, but the most dramatic illustrations are obtained with chirp signals.

The first example is a factor-of-two $(\alpha = 2)$ time-expansion of a sinusoid of constant amplitude whose frequency sweeps linearly from the center frequency of channel 30 $(\Omega_{30} = \frac{2\pi 30}{N})$ to the center frequency of channel 40 in 80 analysis STFT frames. This signal was analyzed with the standard phase-vocoder technique described above. The fast Fourier transform (FFT) size was set to 1024, the analysis hop size was 128, the synthesis hop size was 256, and both analysis and synthesis used Hanning windows of size 1024. Instead of conforming to (10), however, the initial synthesis phases were set equal to the initial analysis phases:

$$\phi_s(0,k) = \angle X(0,\Omega_k)$$

This is a standard initialization choice because it makes it possible to switch from a nonmodified signal $(\alpha = 1)$ to a modified signal without introducing any phase discontinuity.

Fig. 1 shows the amplitude envelope of the resulting signal in the time-domain (obtained by plotting the maximum and minimum values of the sinusoid for each period). The clearly visible amplitude modulation is due to (10) not being satisfied. Fig. 2 shows the analysis and synthesis phases for successive short-time Fourier transform frames. The figure was obtained by measuring the phases at the maximum of the analysis $X(t_a^u, \Omega_k)$ or synthesis $Y(t_s^u, \Omega_k)$ Fourier transforms in each frame, then unwrapping them along successive frames. The analysis phase shows the characteristic parabolic shape due to the linearly varying frequency. The synthesis phase roughly follows this parabolic shape, but exhibits "discontinuities" at frames 38, 53, and (to a lesser extent) 22. These phase jumps result from θ_k not being a constant in (10); it can be easily verified that they occur when the instantaneous frequency of the chirp jumps from one channel to the next. The consistency measure for this resynthesis was $D_M = -10$ dB.

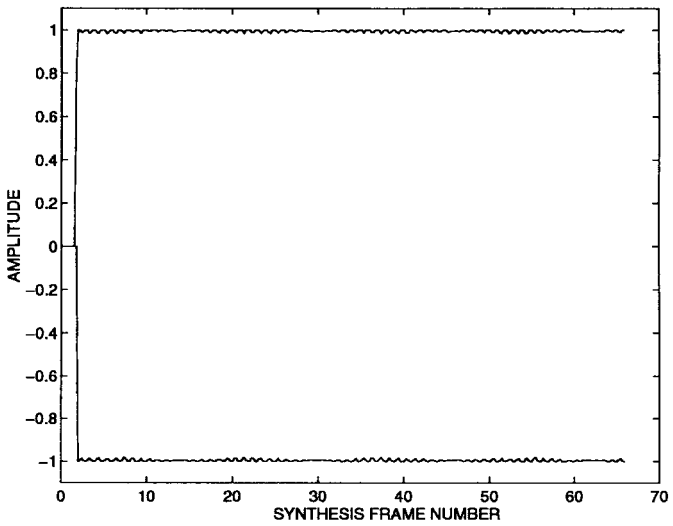

Fig. 3. Factor-two time-scaling of a constant amplitude chirp, "correct"initial phases. Time-domain amplitude-envelope of the modified signal.

If the initial synthesis phases $\phi_s(0,k)$ are set to

$$\phi_s(0,k) = \alpha \angle X(0,\Omega_k) \quad \forall k \tag{11}$$

with $\alpha = 2$, then the time-scaling operation yields the signal shown in Fig. 3. The resulting sinusoid shows only marginal amplitude modulation, and the synthesis phase is free of jumps. This confirms the fact that setting $\theta_k = C$ *for integer modification factors* provides a significant improvement in phase coherence. The consistency in this case was measured to be $D_M = -25$ dB, far better than the preceding result.

When the modification factor is no longer an integer, the resulting signal exhibits phase coherence problems even if the initial phases *are* set according to (11). This can be seen in Fig. 4. The same chirp signal as above is time-scaled by a

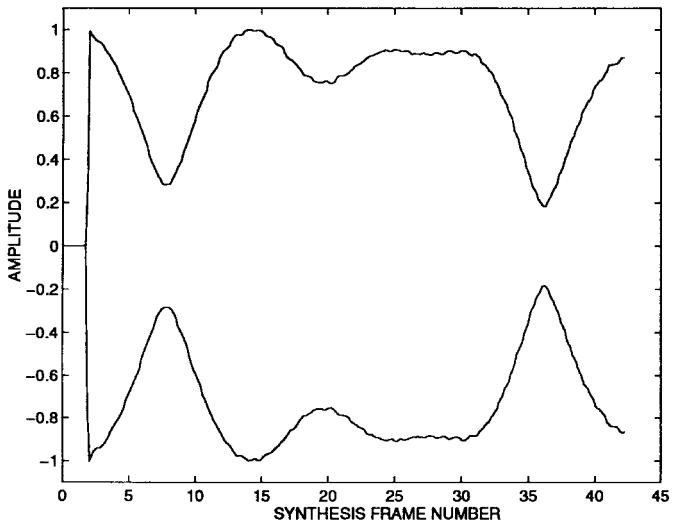

Fig. 4. Factor-1.4 time-scaling of a constant amplitude chirp, initial phases set according to (11). Time-domain amplitude-envelope of the modified signal.

factor $\alpha = 1.4$, with the phase initialization of (11). The resulting signal is severely modulated in amplitude, a result of phase-unwrapping errors in channels distant from channel 30. These errors caused erroneous unwrapping factors to be added to the synthesis phase when the sinusoid's instantaneous frequency reached those channels. The consistency here was measured as $D_M = -6.5$ dB.

Taken together, these very simple test signals demonstrate that, for integer modification factors, initial phases should be set according to (11) to avoid phase jumps. Moreover (and regardless of how the synthesis phases are initialized), time-scale operations with noninteger modification factors are likely to introduce phase discontinuities, leading to significant phase incoherence in the synthesis STFT.

Informal listening tests on speech and music signals confirm these findings. For factor-of-two modifications, high-quality results are obtained when the synthesis phases are initialized according to (11), while phasiness occurs for any other type of initialization. Noninteger modifications always sound phasy, no matter what phase initialization is used.

In addition, time-scaling operations with integer modification factors greater than or equal to three exhibit an increasing level of phasiness even with proper phase initialization. The explanation for this seems to lie in the fact that, for such large modification factors, satisfying (10) still does not guarantee consistency for the synthesis STFT as a whole. Equation (10) ensures vertical phase coherence, but the modified sequence of STFT *magnitudes* may be inconsistent.

III. Old and New Strategies for Reducing Phasiness

A. Magnitude-Only Reconstruction

The phase vocoder is not the only frequency-domain technique that can be applied to the problem of time scaling. Phase unwrapping issues can be avoided entirely by 1) reconstructing purely from the STFT magnitude, or 2) fitting the STFT to an explicit sum-of-sinusoids model.

Algorithms for STFT magnitude-only reconstruction were presented in [4] and [8]. Unfortunately, these algorithms require numerous iterations of the STFT analysis-synthesis cycle in order to arrive at an internally consistent STFT. This makes them far too computationally demanding for contemporary real-time applications. In addition, while convergence has been proven for some of these methods, there is no guarantee that a global minimum will be reached. More recently, various authors have noted that the iterative process can be greatly accelerated by calculating good sets of initial STFT phase values [13]. However, this procedure is still considerably more computation-intensive than the phase vocoder.

The other frequency-domain alternative to the phase vocoder is so-called sinusoidal modeling. This approach is also more computationally demanding than the phase vocoder, and it introduces its own perceptual artifacts which are beyond the scope of this paper.

B. Loose Phase Locking

A suboptimal but less computation-intensive solution to the phase vocoder's phase-unwrapping errors is simply to apply *a-posteriori* constraints to the synthesis phases. An efficient mechanism for accomplishing this was introduced by Puckette [10]. Recognizing that a constant-amplitude, constant-frequency sinusoid in channel k should have identical analysis phases in all nearby channels, Puckette proposed a simple way to loosely constrain synthesis phases to obey the same rule. For each channel in the synthesis STFT, the synthesis phase is calculated by use of the standard phase-propagation formula, (6). However, the final synthesis phase attributed to channel k is that of the complex number

$$Y\left(t_s^u, \Omega_k\right) + Y\left(t_s^u, \Omega_{k-1}\right) + Y\left(t_s^u, \Omega_{k+1}\right)$$

As a result, if channel k is the maximum of the Fourier transform magnitude, its phase is basically unchanged, because $Y(t_s^u, \Omega_{k-1})$ and $Y(t_s^u, \Omega_{k+1})$ are of much lower amplitude. But if channel k is left of the maximum, its phase will be roughly that of $Y(t_s^u, \Omega_{k+1})$ whose magnitude is larger than $Y(t_s^u, \Omega_k)$: the channels around a peak in the Fourier transform are "phase locked" to the peak.

This technique is attractive, especially because it requires only a few additional multiplications per channel. It produces a visible improvement in the phase vocoder output for simple test signals and a measurable improvement in STFT consistency. However, informal listening tests show that the reduction in phasiness is very signal-dependent and, unfortunately, never dramatic.

C. Rigid Phase Locking

The fundamental limitation (and also the attraction) of the loose phase locking scheme is that it avoids any explicit determination of the signal structure: the same calculation is performed in every channel, independently of its content. The result is that synthesis phases in the channels around a given sinusoid only gradually and approximately develop vertical phase coherence. If we *really* want to restore vertical phase coherence to the synthesis STFT, we need to take a step closer

to an actual sum-of-sinusoids model. We now present two versions of a new phase-locking technique, inspired by the loose phase locking above, but based on the explicit identification of peaks (and thus, presumably, sinusoids) in the spectrum.

The new phase-updating technique begins with a coarse peak-picking stage where vocoder channels are searched for local maxima. In the simplest implementation, a channel whose amplitude is larger than its four nearest neighbors is said to be a peak; this criterion is both simple and cost-effective (although admittedly primitive). The series of peaks subdivides the frequency axis into "regions of influence" located around each peak. The basic idea is to update the phases *for the peak channels only* according to the standard phase-propagation (6); the phases of the remaining channels within each region are then "locked" in some way to the phase of the peak channel. In our experiments, the upper limit of the region around peak Ω_{k_l} was set to the middle frequency between that peak and the next one $(\Omega_{k_l} + \Omega_{k_{l+1}})/2$. Another reasonable choice would be the channel of lowest amplitude between the two peaks.

1) Identity Phase Locking Rather than imposing a strong phase-equality constraint (based on the assumption that the peak is the trace of a constant-amplitude, constant-frequency sinusoid), one can constrain the synthesis phases around the peak to be related in the same way as the analysis phases: the phase differences between successive channels around a peak are made identical in the synthesis STFT to the corresponding phase differences in the analysis Fourier transform. If Ω_{k_l} is the center frequency of the dominant peak, we set

$$\angle Y\left(t_s^u, \Omega_k\right) = \angle Y\left(t_s^u, \Omega_{k_l}\right) + \angle X\left(t_a^u, \Omega_k\right) - \angle X\left(t_a^u, \Omega_{k_l}\right) \tag{12}$$

for all channel k in the peak's region of influence, as defined above.

Note that the idea of preserving the phase-relations between nearby bins has been independently proposed in [2] in the context of integer-factor time-scale modifications, and previously in [11] as a means to reduce transient smearing.

This phase-locking mechanism significantly improves the consistency of the resulting series of STFT and greatly reduces the phasiness of the modified signal. Applying this technique to the chirp signal of our previous example (again with a time-scale factor of 1.4) produces the signal shown in Fig. 5.

The resulting signal shows no sign of amplitude modulation, and the consistency distance was measured to be $D_M = -37$ dB, a very large improvement over the mere -6.5 dB of the non phase-locked modification.

Identity phase locking has two major computational advantages. First, since phase unwrapping is performed only on peak channels, the instantaneous frequency of the underlying sinusoid is sure to be near the center frequency of the channel in question. This means that the phase-unwrapping constraint $R_a\omega_h < \pi$ can be relaxed, and larger values of R_a can be used. In practice, an input overlap of 50% is possible without generating phase-unwrapping errors. Since the usual requirement for phase-vocoder-based time-scale modification is a minimum of 75% overlap (for standard analysis windows, such as Hanning or Hamming), identity phase locking essentially cuts the computational cost in half!

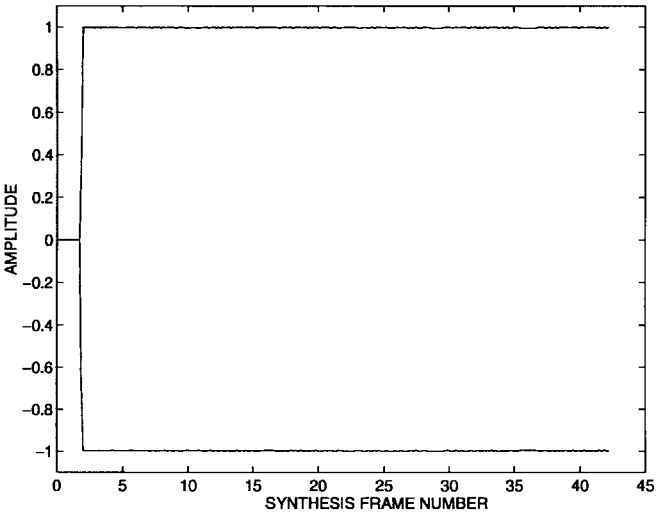

Fig. 5. Factor-1.4 time-scaling of a constant amplitude chirp, identity peak-phase-locking. Time-domain amplitude-envelope of the modified signal.

Second, this new technique requires trigonometric calculations *only for peak channels*: once the synthesis phase of the peak channel has been determined, one can calculate the angle θ required to rotate $X(t_a^u, \Omega_{k_l})$ into $Y(t_s^u, \Omega_{k_l})$ as follows:

$$\theta = \angle Y\left(t_s^u, \Omega_{k_l}\right) - \angle X\left(t_a^u, \Omega_{k_l}\right) \tag{13}$$

then calculate the phasor $Z = e^{j\theta}$ and obtain the neighboring channels by use of simple complex algebra

$$Y\left(t_s^u, \Omega_k\right) = ZX\left(t_a^u, \Omega_k\right) \tag{14}$$

which can be easily shown to satisfy the phase-locking equation (12): *neighboring channels only require one complex multiply*!

The identity phase-locking scheme can be summarized in the following steps.

1) For the new STFT frame, locate prominent peaks.
2) For each peak, calculate the instantaneous frequency using horizontal phase unwrapping, and calculate the updated synthesis phase, according to (6).
3) Calculate rotation angle θ according to (13) and phasor $Z = e^{j\theta}$.
4) Apply rotation to all channels around and including peak channel, according to (14).
5) Repeat the above steps for the next peak, until all peaks have been processed.
6) Proceed to the next synthesis frame.

2) Scaled Phase Locking: An improvement over the preceding technique comes from recognizing that if a peak switches from channel k_0 at frame $u-1$ to channel k_1 at frame u the unwrapping equation (5) should be based on $\angle X(t_a^u, \Omega_{k_1}) - \angle X(t_a^{u-1}, \Omega_{k_0})$ instead of $\angle X(t_a^u, \Omega_{k_1}) - \angle X(t_a^{u-1}, \Omega_{k_1})$. Likewise, the phase-propagation equation (6) should be

$$\angle Y\left(t_s^u, \Omega_{k_1}\right) = \angle Y\left(t_s^{u-1}, \Omega_{k_0}\right) + R_s\hat{\omega}_{k_1}\left(t_a^u\right) \tag{15}$$

where the phase increment $R_s\hat{\omega}_{k_1}(t_a^u)$ is accumulated to $\angle Y(t_s^{u-1}, \Omega_{k_0})$ rather than $\angle Y(t_s^{u-1}, \Omega_{k_1})$. It is easy to show

that in that case, the synthesis phase at the peak channel corresponding to sinusoid i at time t_a^u is indeed $C+\alpha\phi_i(t_a^u)$, which is not necessarily the case in the preceding technique.

The problem is then to determine what peak in frame $u-1$ corresponds to the peak Ω_{k_1} in frame u. A very simple way of doing this is to pick the peak of the region to which channel Ω_{k_1} belonged in frame $u-1$. Accordingly, to calculate the synthesis phase of channel k_1 at frame u, one can simply look up the dominant peak in the region channel k_1 belonged to in frame $u-1$, and use its analysis and synthesis phases, when applying (5) and (6). This being done, the neighboring channels can be synchronized to the peak, and the identity phase-locking equation can be generalized as

$$\angle Y\left(t_s^u,\Omega_k\right) = \angle Y\left(t_s^u,\Omega_{k_l}\right)+\beta\left[\angle X\left(t_a^u,\Omega_k\right)-\angle X\left(t_a^u,\Omega_{k_l}\right)\right] \quad (16)$$

where β is a phase scaling factor. Identity phase locking is simply $\beta=1$. Exactly how the phases should be modified upon synthesis to ensure proper vertical phase coherence is not easy to assess, as explained in Section II-B1. However, it appears that identity phase locking can be further improved by setting β to a value between one and α. There is little theoretical ground to justify such a choice, since the phases in the modified signal should probably not be related to those in the original signal in a linear way. However, when integer modification factors α are used in the standard implementation of the phase-vocoder, and when the initialization of (11) is used, it is easy to verify that phase-differences are also scaled by $\beta=\alpha$. Moreover, informal listening tests have shown that setting $\beta\approx 2/3+\alpha/3$ helps further reduce phasiness. Note that the phases $\angle X(t_a^u,\Omega_k)$ must be unwrapped across channels k around the peak channel before applying (16), in order to avoid $2\beta\pi$ channel jumps in the synthesis phases.

In contrast to identity phase locking, scaled phase locking does not permit the STFT values in neighboring channels to be calculated simply via a complex multiplication; therefore its implementation requires somewhat more computation. On the other hand, the quality of the time-scaled signals has been found to be consistently higher with scaled phase-locking than with identity phase locking.

The scaled-phase-locking scheme can be summarized in the following steps.

1) For the new STFT frame, locate prominent peaks.
2) For each peak channel k_i, locate corresponding peak in preceding frame, calculate instantaneous frequency using horizontal phase unwrapping, and calculate updated synthesis phase according to (15).
3) Unwrap analysis phases across all channels in the region of influence.
4) For each channel around the peak channel, calculate analysis phase difference between peak and current channel, and calculate current synthesis phase using (16).
5) Repeat the above steps for the next peak, until all peaks have been processed.
6) Proceed to the next synthesis frame.

TABLE I
CONSISTENCY MEASURE FOR VARIOUS SYNCHRONIZATION TECHNIQUES, CHIRP SIGNAL

Synchronization type	D_M
None	+1.5 dB
Loose phase locking	-13 dB
Iterative reconstruction, 5 iterations	-16 dB
Iterative reconstruction, 50 iterations	-19 dB
Identity phase locking	-30 dB
Scaled phase locking	-30 dB

TABLE II
CONSISTENCY MEASURE FOR VARIOUS SYNCHRONIZATION TECHNIQUES, SPEECH SIGNAL

Synchronization type	D_M
None	0 dB
Loose phase locking	-11 dB
Iterative reconstruction, 5 iterations	-14 dB
Iterative reconstruction, 50 iterations	-18 dB
Identity phase locking	-15 dB
Scaled phase locking	-14 dB
PSOLA	-9 dB

D. Comparison of Phase-Locking Techniques

1) Consistency Measure: Although the consistency measure of (7) is not a very accurate measure of phasiness, it still gives an indication of how well various synchronizing schemes perform. Two signals were used to compare the synchronizing techniques: a chirp sinusoid identical to that used in Section II-C4, and a segment of speech signal. The speech signal was a male speaker uttering the word "before," sampled at 16 kHz.

Four synchronization techniques were compared, along with the standard, nonsynchronized phase-vocoder technique. The synchronization techniques used were Puckette's loose phase locking [10], iterative magnitude-only reconstruction [4] (measured after five and 50 iterations), identity phase locking, and scaled phase locking with $\beta=1.4$. The modification factor was $\alpha=2.2$, and the initial synthesis phases were set to the initial analysis phases $\phi_s(0,k)=\angle X(0,\Omega_k)$. The FFT size was 1024, with an output hop factor of 256.

Table I presents the consistency measure for the chirp signal. The results indicate that some type of synchronization is required to insure some degree of consistency. As expected, loose phase locking is significantly less effective than either identity or scaled phase-locking. Iterative magnitude-only reconstruction, even after 50 iterations, still does not match our synchronization techniques as the iterative procedure seems to be caught in a local minimum. Most of the consistency improvement is achieved within the first few iterations.

Table II presents the consistency measure for the speech signal. Again, our phase-synchronization techniques outperform loose phase locking, but some marginal improvement in consistency can be obtained by the iterative procedure, with a large number of iterations. Also included is the consistency measure for pitch-synchronous overlap add (PSOLA), a high-quality time-domain technique based on overlap-adding small segments of waveform [7].

2) Informal Listening Tests Listening tests on speech and polyphonic musical signals partially confirm the results of the

consistency measure. The standard, nonsynchronized phase-vocoder technique has the poorest consistency measure, and it always sounds worse than the other techniques.

However, the consistency measure does not always accurately reflect the perceived phasiness of the modified signal. For example, identity phase locking and scaled phase locking have similar consistency scores, but the latter consistently sounds better than the former. Conversely, the higher consistency measure obtained by the iterative reconstruction does not correspond to any perceptible quality improvement. In fact, the modified signal is plagued with an undesirable roughness that other techniques do not exhibit (an artifact mentioned in the original paper [4]). Nevertheless, it appears that low values of the consistency measure (above -5 dB) always indicate the presence of phasiness in the signal.

The influence of the parameter β in scaled phase locking is more dramatic when 75% overlap is used. In that case, and for larger modification factors ($\alpha > 2$), identity phase locking can still be somewhat phasy. Using $\beta = \alpha$ in (16) significantly reduces phasiness. When 50% overlap is used, however, β must be kept closer to one to avoid undesirable roughness in the output signal. On speech signals, the modified voice has more presence, and nonvoiced segments sound slightly more natural with scaled phase locking than with identity phase locking.

For integer modification factors, the standard algorithm *without phase unwrapping* but with proper initial conditions (such that $\theta_k = C\ \forall k$) yields high-quality results which only scaled phase locking can match for noninteger modification factors. Some residual phasiness can still be heard in the modified signals, especially for larger modification factors ($\alpha > 3$).

Finally, when speech signals are processed, all the above phase-locked techniques still exhibit more reverberation or phasiness than time-domain techniques such as the PSOLA technique [7]. Surprisingly, the consistency measure for the speech signal modified via the PSOLA technique is worse than with other techniques; this confirms that the consistency measure is not a clear indicator of phasiness.

E. Conclusion

We have presented a detailed explanation for the presence of phasiness or reverberation in signals that have been time scaled by phase-vocoder techniques. We have shown that sinusoidal components crossing channels generate both phase incoherence between adjacent channels and phase "discontinuities" between successive frames. For integer modification factors, a suitable choice of initial analysis and synthesis phases eliminates this problem entirely, yielding high-quality results. For noninteger modification factors, phase incoherence can be avoided by using one of the two phase-locking techniques presented above. These techniques dramatically reduce the phasiness of the output signal and also enable the use of 50% overlap between frames; this decreases the computational cost by at least a factor of two.

Although the quality of the modified signals obtained via phase locking is far better than that obtained using the standard phase-vocoder technique, it nevertheless remains inferior to the quality attainable via time-domain techniques for monophonic, pitched signals such as speech. One reason for this is that, ideally, not only phases but also amplitudes should be modified when performing time-scale modification. In the frequency domain, a chirp sinusoid has a wider center lobe than a constant-frequency sinusoid. Time-stretching it by a large factor should turn it into a near-constant frequency with a narrower center lobe. This fact is not taken into account in the phase-vocoder techniques presented in this paper. Incorporating such refinements could increase the quality of the modification, but the resulting complication might rival that of sinusoidal-modeling methods [6]. Refer to [11] for the description of a technique which attempts to modify both STFT phases and amplitudes.

Another reason might be linked to the lack of phase-synchronization between harmonics of the same fundamental, or lack of "shape-invariance" as it has been called in the literature. Achieving shape-invariance would require synchronizing peak-phases with one another, an issue which is not addressed by the techniques presented here. Also, note that the concept of shape-invariance is only valid for sounds made up of quasiharmonic signals. Refer to [3], [12], or [14] for the description of techniques that attempt to achieve shape-invariance.

ACKNOWLEDGMENT

The authors would like to thank O. Cappé for his help in reviewing early versions of this contribution, and the reviewers for their insightful suggestions to improve the clarity of this paper.

REFERENCES

[1] M. Dolson, "The phase vocoder: A tutorial," *Comput. Music J.*, vol. 10, pp. 14–27, 1986.
[2] H. J. S. Ferreira, "An odd-DFT based approach to time-scale expansion of audio signals," to be published.
[3] E. B. George and M. J. T. Smith, "Analysis-by-synthesis/Overlap-add sinusoidal modeling applied to the analysis and synthesis of musical tones," *J. Audio Eng. Soc.*, vol. 40, pp. 497–516, 1992.
[4] D. W. Griffin and J. S. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-32, pp. 236–243, Apr. 1984.
[5] J. Laroche, "Time and pitch scale modification of audio signals," in *Applications of Digital Signal Processing to Audio and Acoustics*, M. Kahrs and K. Brandenburg, Eds. Boston, MA: Kluwer, 1998.
[6] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-34, pp. 744–754, Aug. 1986.
[7] E. Moulines and J. Laroche, "Non parametric techniques for pitch-scale and time-scale modification of speech," *Speech Commun.*, vol. 16, pp. 175–205, Feb. 1995.
[8] S. H. Nawab, T. Quatieri, and J. S. Lim, "Signal reconstruction from short-time fourier transform magnitude," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-31, pp. 986–998, Aug. 1983.
[9] R. Portnoff, "Time-scale modifications of speech based on short-time Fourier analysis," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 29, pp. 374–390, 1981.
[10] M. S. Puckette, "Phase-locked vocoder," *IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, 1995.
[11] T. F. Quatieri, R. B. Dunn, and T. E. Hanna, "A subband approach to time-scale expansion of complex acoustic signals," *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 515–519, Nov. 1995.
[12] T. F. Quatieri and J. McAulay, "Shape invariant time-scale and pitch modification of speech," *IEEE Trans. Signal Processing.*, vol. 40, pp. 497–510, Mar. 1992.

[13] S. Roucos and A. M. Wilgus, "High quality time-scale modification of speech," in *Proc. IEEE ICASSP'85*, Tampa, FL, pp. 493–496.
[14] B. Sylvestre and P. Kabal, "Time-scale modification of speech using an incremental time-frequency approach with waveform structure compensation," in *Proc. IEEE ICASSP'92*, pp. 81–84.

Jean Laroche (M'92) was born in Bordeaux, France, on July 20, 1963. He received the M.S. degree from the Ecole Polytechnique in 1986 and the Ph.D. degree from the Ecole Nationale Supérieure des Télécommunications, Paris, France, in 1989, both in signal processing.

In 1990 he was a Fellow of the ITT international grant at the Center for Music Experiment, University of California, San Diego. In 1991, he became an Assistant Professor at Telecom Paris in the Signal Department, teaching audio and speech processing and acoustics. Since 1996, he has been a researcher at the Joint E-mu/Creative Technology Center, Scotts Valley, CA, helping design techniques for advanced music and audio processing.

Mark Dolson (S'81–M'82) was born in Pittsburgh, PA, in 1955. He received the B.S. degree in electrical engineering from Carnegie Mellon University, Pittsburgh, PA, in 1976, and the Ph.D. degree in electrical engineering from California Institute of Technology, Pasadena, in 1983.

From 1978 through 1982, he was a Scientist at Hughes Research Laboratory, Malibu, CA. He spent the next 11 years as a Research Scientist at the University of California, San Diego, focusing on the application of digital signal processing to problems in music, speech, and perception. In 1994, he joined the Joint E-mu/Creative Technology Center, Scotts Valley, CA, where he is now Manager of Digital Signal Processing Research.