Speech Analysis/Synthesis and Modification Using an Analysis-by-Synthesis/Overlap-Add Sinusoidal Model

E. Bryan George, *Member, IEEE,* and Mark J. T. Smith, *Fellow, IEEE*

***Abstract*— Sinusoidal modeling has been successfully applied to a broad range of speech processing problems, and offers advantages over linear predictive modeling and the short-time Fourier transform for speech analysis/synthesis and modification. This paper presents a novel speech analysis/synthesis system based on the combination of an overlap-add sinusoidal model with an analysis-by-synthesis technique to determine model parameters. The paper describes this analysis procedure in detail, and introduces an equivalent frequency-domain algorithm that takes advantage of the computational efficiency of the fast Fourier transform (FFT). In addition, a refined overlap-add sinusoidal model capable of shape-invariant speech modification is derived, and a pitch-scale modification algorithm is defined that preserves speech bandwidth and eliminates noise migration effects. Analysis-by-synthesis achieves very high synthetic speech quality by accurately estimating component frequencies, eliminating sidelobe interference effects, and effectively dealing with nonstationary speech events. The refined overlap-add synthesis model correlates well with analysis-by-synthesis, and modifies speech without objectionable artifacts by explicitly controlling shape invariance and phase coherence. The proposed analysis-by-synthesis/overlap-add (ABS/OLA) system allows for both fixed and time-varying time-, frequency-, and pitch-scale modifications, and computational shortcuts using the FFT algorithm make its implementation feasible using currently available hardware.**

***Index Terms*— Sinusoidal modeling, speech analysis, speech enhancement, speech synthesis.**

I. Introduction

GENERALLY defined, speech modification is the process of changing certain perceptual properties of speech while leaving other properties unchanged. Among the many types of speech information that may be altered are rate of articulation, bandwidth, pitch, message content, and formant characteristics. Speech modification techniques are used in a variety of applications related to speech communication. For instance, time-scale expansion is used to slow rapid or degraded speech, enhancing its intelligibility [1]. Conversely, time-scale compression is used in message playback systems for fast scanning of recorded messages. Frequency-scale modification is often performed to transmit speech over limited bandwidth communication channels [2] or to place speech in a desired bandwidth as an aid to the hearing impaired [3]. Pitch-scale modification, which changes the fundamental frequency of speech while maintaining the original formant structure, is useful in text-to-speech systems based on concatenation of speech segments [4]. Conversely, formant modification techniques are used to compensate for vocal tract defects and helium speech environments.

A variety of model-based approaches to speech modification have been discussed in the literature: The classical pitch-excited linear predictive coding (LPC) model [5], which represents the speech production process in terms of a spectrally flat excitation signal driving a slowly-varying vocal tract filter, is capable of modifying analyzed speech in a number of useful ways and is the basis of many text-to-speech systems. However, pitch-excited LPC is sensitive to errors in pitch and voicing state estimation, does not work well for certain speakers or for signals that fail to fit the given "source/filter" model, and performs poorly in the presence of background noise.

Speech modification techniques based on time-frequency representations, particularly the discrete short-time Fourier transform (DSTFT), exploit the observation that speech signals are quasiperiodic, short-time stationary sequences [1], [6]. The digital phase vocoder (DPV) [7] relates the amplitudes and frequencies of the outputs of a digital filterbank to excitation and vocal tract properties. The refined phase vocoder proposed by Portnoff [8] takes advantage of the computational efficiency of the fast Fourier transform (FFT) for implementation. However, phase distortions in the presence of modification cause dispersion and shape variation in the modified speech signal, resulting in synthetic speech with an objectionable reverberant quality. Griffin and Lim have developed an iterative signal estimation algorithm that enhances the quality of modified speech compared to Portnoff's method [9], but the technique requires a considerable amount of added computation.

The difficulties of speech modification using the DSTFT arise largely from the use of a harmonic sinusoidal model to represent quasiharmonic speech signals, and from the rather indirect relationship between model parameters and fundamental speech properties. In an attempt to deal with these issues, a less constrained time–frequency model of speech was

Manuscript received May 20, 1993; revised September 20, 1996. This work was sponsored by the NSF under Contract DCI-8611372. The associate editor coordinating the review of this paper and approving it for publication was Dr. Daniel Kahn.

E. B. George was with the Signal Processing Center of Technology, Lockheed-Martin, Inc., Nashua, NH 03061 USA. He is now with the DSP Solutions Research and Development Center, Texas Instruments Incorporated, Dallas, TX 75265 USA (e-mail: ebg@ti.com).

M. J. T. Smith is with the School of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: mjts@eedsp.gatech.edu).

Publisher Item Identifier S 1063-6676(97)06384-0.

applied to low-rate speech coding by Almeida and Tribolet [10], who developed a *generalized harmonic* representation of speech to account for short-term pitch and vocal tract variations. Hedelin [11] also generalized sinusoidal modeling in the context of speech coding, by representing speech as a sum of amplitude- and frequency-modulated sinusoids whose time-varying behavior directly reflects the voicing state, pitch, and formant structure of speech.

Sinusoidal speech modeling as applied to speech analysis/synthesis and modification was further developed by McAulay and Quatieri [12], [13]. Their work introduced a variety of techniques for dealing with the estimation and modeling problems encountered in sinusoidal speech modeling, resulting in the well-known *sinusoidal transformation system* (STS). In the STS, a given speech signal is represented as a sum of sinusoids with continuously variable amplitude and frequency tracks. Analysis of model parameters is carried out over short frames assuming short-time stationarity; under this assumption, frames of speech are modeled as a sum of constant-amplitude sinusoids with constant (but arbitrary) frequencies. The parameters of these sinusoids are determined at fixed time intervals by picking the peaks of a finely sampled discrete Fourier transform (DFT) of the windowed speech frame.

Given analyzed sinusoidal model parameters, a set of continuously variable parameter tracks representing the speech signal is constructed by matching nearest neighbor components from one frame to the next and interpolating the matched parameters over time using polynomial functions. The functional form of the resulting parameter set permits time- and frequency-scale modification by relatively straightforward operations on the parameter tracks [13]. McAulay and Quatieri have formulated a technique referred to as *shape invariant modification* [14], [15]. This technique substantially reduces the reverberance associated with DSTFT speech modification by attempting to preserve the original waveform shape in modified speech. To reduce synthesis computation in the STS, an overlap-add model using the inverse FFT algorithm has also been proposed [16].

This paper presents a novel sinusoidal speech modeling system applied to speech analysis/synthesis and modification [17]–[19]. In contrast to the STS, the proposed analysis/synthesis system uses a successive approximation-based analysis-by-synthesis procedure rather than peak-picking to determine model parameters. Analysis-by-synthesis, which has been successfully applied to both time–frequency [20], [21] and linear predictive [22] speech coders, is well suited to the nonlinear nature of sinusoidal model analysis and enhances modeling accuracy. Another distinction of the proposed system is its use of overlap-add sinusoidal modeling to affect speech modifications directly. Overlap-add synthesis is well matched to analysis-by-synthesis and, as mentioned above, is computationally efficient. Furthermore, the paper demonstrates that a refined quasiharmonic overlap-add model formulation designed for shape invariant modification can produce synthetic speech without objectionable artifacts.

The resulting *analysis-by-synthesis/overlap-add* (ABS/OLA) system is capable of fixed and time-varying time- and frequency-scale modification, and allows for joint time and frequency modifications as well. In addition, the paper describes an approach to pitch-scale modification that deals with the problems of bandwidth compression and noise amplification, two common difficulties encountered in pitch-scaling systems. When combined with techniques to reduce the computation of analysis and synthesis, the ABS/OLA system provides a combination of synthetic speech quality and reasonable computational load not seen in other high-quality modification systems.

The paper is organized as follows: Section II introduces the sinusoidal model formulation used for speech analysis/synthesis, reviews analysis-by-synthesis techniques described in [17]–[19] to determine model parameters, and formulates analysis-by-synthesis as a dual frequency-domain algorithm. Section III describes the refined overlap-add sinusoidal model required for speech modification, and develops a straightforward framework for its application to time-, frequency-, and pitch-scale modification of speech. Section IV relates frequency-domain analysis-by-synthesis to the fast Fourier transform algorithm, providing a significant reduction of the required computational load of analysis, and discusses techniques to implement the refined overlap-add model using the inverse FFT algorithm. Section V discusses the results of comparisons of the ABS/OLA and STS systems, and suggests areas of future research. Finally, Section VI concludes with a brief summary.

II. Synthesis Model and Analysis Procedure

This section proposes a basic overlap-add sinusoidal model designed to represent speech signals accurately, and reviews the analysis-by-synthesis approach to analyzing sinusoidal model parameters discussed in [17]–[19].

A. Overlap-Add Sinusoidal Model

The model proposed to represent a real-valued, uniformly sampled speech signal $s[n]$ is a variation on the classic overlap-add sinusoidal model formulation which is given in its most general form by

$$\tilde{s}[n] = \sigma[n] \sum_{k=-\infty}^{\infty} w_s[n - kN_s]\tilde{s}^k[n - kN_s]. \qquad (1)$$

The synthesis window $w_s[n]$ is a complementary window[1] obeying the constraint

$$\sum_{k=-\infty}^{\infty} w_s[n - kN_s] = 1$$

for all n, where N_s determines the separation between adjacent synthesis windows. The interval $kN_s \leq n < (k+1)N_s$ is defined as the k-th *synthesis frame*, thus N_s may be referred to unambiguously as the synthesis frame length. The kth *synthetic contribution* sequence $\tilde{s}^k[n]$ is a sum of constant-amplitude,

[1] Typically a symmetric, tapered window such as a triangular or Hanning window.

constant-frequency sinsusoids given by

$$\tilde{s}^k[n] = \sum_{j=1}^{J[k]} A_j^k \cos(2\pi f_j^k n/F_s + \phi_j^k) = \sum_{j=1}^{J[k]} A_j^k \cos(\omega_j^k n + \phi_j^k) \quad (2)$$

where F_s is the sampling frequency of $s[n]$, and where $0 \leq f_j^k \leq F_s/2$. Given a symmetric synthesis window, $\tilde{s}^k[n]$ has a region of support that is also symmetric about $n = 0$. The modulating "envelope sequence" $\sigma[n]$ is a moving weighted average of the magnitude of $s[n]$, computed over a number of synthesis frames. The purpose of $\sigma[n]$ is to provide a concise representation of syllabic energy variations in $s[n]$, and to reduce the effects of these variations on parameter estimation. In particular, using $\sigma[n]$ to model global amplitude modulation avoids the need for extraneous sinusoidal components to model this behavior [17].

This sinusoidal model formulation resembles overlap-add synthesis using the DSTFT in that constant-amplitude, constant-frequency sinusoids represent $s[n]$ on a frame-by-frame basis, but differs in its use of the modulating envelope sequence $\sigma[n]$, variable numbers of sinusoidal components, and arbitrary component frequencies. Given a symmetric synthesis window of length $2N_s + 1$, a synthesis frame of N_s samples of $\tilde{s}[n]$ may be expressed in the following relatively compact form:

$$\tilde{s}[n + kN_s] = \sigma[n + kN_s](w_s[n]\tilde{s}^k[n] + w_s[n - N_s]\tilde{s}^{k+1}[n - N_s]) \quad (3)$$

for $0 \leq n < N_s$. As with any frame-based approach to speech modeling, care must be taken in choosing N_s such that the speech signal may be assumed stationary over a given frame interval [23]. Typical values of N_s for speech signals correspond to between 5 and 20 ms.

B. Analysis-by-Synthesis

The parameter set that must be determined in order to represent $s[n]$ consists of the envelope sequence $\sigma[n]$ and the amplitudes $\{A_j^k\}$, frequencies $\{\omega_j^k\}$, and phases $\{\phi_j^k\}$ of each synthetic contribution $\tilde{s}^k[n]$. The purpose of the envelope sequence $\sigma[n]$ is to represent syllabic variations in the average magnitude of $s[n]$ over an analysis frame. Since such variations are inherently low bandwidth, $\sigma[n]$ can be reasonably estimated by lowpass filtering $|s[n]|$. To optimize modeling accuracy and robustness, $\sigma[n]$ should be able to accurately track transient speech energy variations but should not contain any speech components that introduce periodic ripple for voiced speech. A recursive, quasi-Gaussian lowpass filter design that exhibits a good tradeoff between these requirements is described in [17]–[19], although any reasonable lowpass filter with an approximate cutoff frequency less than 40 Hz may be used. Model accuracy in the ABS/OLA system is not especially sensitive to the estimation of $\sigma[n]$, in the sense that any reasonable estimate is likely to improve model accuracy.

Given $\sigma[n]$, the objective of analysis is to determine amplitude, frequency, and phase parameters for each $\tilde{s}^k[n]$ in (1) such that $\tilde{s}[n]$ is "closest" to $s[n]$ in some sense. An approach often employed to solve problems of this type is to minimize the mean-square modeling error

$$E = \sum_{n=-\infty}^{\infty} \{s[n] - \tilde{s}[n]\}^2$$

in terms of the parameters of $\tilde{s}[n]$. Of course, attempting to solve this problem simultaneously for all the parameters of $\tilde{s}[n]$ is not practical.

Fortunately, if $s[n]$ is approximately stationary over short time intervals, it is feasible to solve for the amplitude, frequency, and phase parameters of each $\tilde{s}^k[n]$ in isolation by approximating a segment of $s[n]$ over an analysis frame of length $2N_a + 1$ samples centered at $n = kN_s$. The synthetic contribution $\tilde{s}^k[n]$ may then be determined by minimizing

$$E^k = \sum_{n=-N_a}^{N_a} w_a[n]\{s[n + kN_s] - \sigma[n + kN_s]\tilde{s}^k[n]\}^2 \quad (4)$$

with respect to the amplitudes, frequencies, and phases of $\tilde{s}^k[n]$. The analysis window $w_a[n]$ may be an arbitrary positive sequence, but is typically a symmetric, finite-length tapered window. Strategies for choosing an analysis window and appropriate values of N_a will be discussed later, but in order to ensure the accuracy of $\tilde{s}[n]$, N_a should be greater than or equal to N_s. Comparing (4) to (3), we note that the symmetric analysis interval $[-N_a, N_a]$ with $N_a \geq N_s$ implies that the analysis region may extend over several synthesis frames.

Unfortunately, without *a priori* knowledge of the frequency parameters, direct minimization of E^k is a highly nonlinear problem that is very difficult to solve. As an alternative, a slightly suboptimal but relatively efficient analysis-by-synthesis algorithm may be employed to determine sinusoidal model parameters using successive approximation. The proposed analysis-by-synthesis algorithm works as follows: We first define "component sequences" $\hat{s}_j^k[n]$ as

$$\hat{s}_j^k[n] \triangleq \sigma[n + kN_s]A_j^k \cos(\omega_j^k n + \phi_j^k). \quad (5)$$

Suppose now that the parameters of $\ell - 1$ components have been determined previously, generating a successive approximation to $s[n]$ in the range $kN_s - N_a \leq n \leq kN_s + N_a$ expressed as

$$\tilde{s}_{\ell-1}^k[n] = \sum_{j=1}^{\ell-1} \hat{s}_j^k[n] = \sigma[n + kN_s] \sum_{j=1}^{\ell-1} A_j^k \cos(\omega_j^k n + \phi_j^k)$$

and a *successive error sequence*

$$e_{\ell-1}^k[n] = s[n + kN_s] - \tilde{s}_{\ell-1}^k[n].$$

Given the initial conditions $\tilde{s}_0^k[n] = 0$ and $e_0^k[n] = s[n + kN_s]$, these sequences may be updated recursively by

$$\tilde{s}_\ell^k[n] = \tilde{s}_{\ell-1}^k[n] + \sigma[n + kN_s]A_\ell^k \cos(\omega_\ell^k n + \phi_\ell^k)$$
$$e_\ell^k[n] = e_{\ell-1}^k[n] - \sigma[n + kN_s]A_\ell^k \cos(\omega_\ell^k n + \phi_\ell^k) \quad (6)$$

for $\ell \geq 1$.

The goal of analysis-by-synthesis is to update the approximation to $s[n]$ by adding a single component such that the updated approximation is as good as possible. This is achieved by minimizing the *successive error norm* E_ℓ^k of $e_\ell^k[n]$, which is given by

$$\begin{aligned} E_\ell^k &= \sum_{n=-N_a}^{N_a} w_a[n]\{e_\ell^k[n]\}^2 \\ &= \sum_{n=-N_a}^{N_a} w_a[n]\{e_{\ell-1}^k[n] - \sigma[n+kN_s] \\ &\quad \times A_\ell^k \cos(\omega_\ell^k n + \phi_\ell^k)\}^2 \end{aligned} \tag{7}$$

in terms of the amplitude, frequency, and phase of the new component. This error formulation makes two points clear. First, each new component models the error remaining after approximation of $s[n+kN_s]$ by previous components, emphasizing the successive nature of the approximation. Second, the analysis window $w_a[n]$ serves to emphasize the errors occuring at different values of n; a tapered, symmetric window centered at $n=0$ is useful since it serves to emphasize the region at which the synthetic contribution $\tilde{s}^k[n]$ dominates the overall approximation to $s[n+kN_s]$ given by (1).

Although the approximation problem has been greatly simplified at this point, it is still not practical to solve simultaneously for the parameters of the new component due to the embedded frequency and phase terms. However, assuming for the moment that ω_ℓ is fixed and expressing the sinusoid as $A_\ell \cos(\omega_\ell n+\phi_\ell) = a_\ell \cos \omega_\ell n + b_\ell \sin \omega_\ell n$ (suppressing frame notation), the problem of approximating $e_{\ell-1}[n]$ becomes a linear least-squares approximation that has the normal equations

$$\begin{aligned} a_\ell\gamma_{11} + b_\ell\gamma_{12} &= \psi_1 \\ a_\ell\gamma_{12} + b_\ell\gamma_{22} &= \psi_2 \end{aligned} \tag{8}$$

where a_ℓ and b_ℓ are the *quadrature parameters* of the new component, and where

$$\begin{aligned} \gamma_{ij} &= \sum_{n=-N_a}^{N_a} w_a[n]\sigma^2[n+kN_s] \cos(\omega_\ell n - (i-1)\pi/2) \\ &\quad \times \cos(\omega_\ell n - (j-1)\pi/2) \\ \psi_i &= \sum_{n=-N_a}^{N_a} w_a[n]e_{\ell-1}[n]\sigma[n+kN_s] \\ &\quad \times \cos(\omega_\ell n - (i-1)\pi/2) \end{aligned} \tag{9}$$

for $1 \le i,j \le 2$. Solving (8) for a_ℓ and b_ℓ leads to closed-form expressions for A_ℓ, ϕ_ℓ, and E_ℓ in terms of the quadrature parameters [17]–[19].

This establishes a method for determining the optimal amplitude and phase parameters for a single sinusoidal component of $\tilde{s}^k[n]$ at a given frequency. To determine an appropriate frequency for this sinusoid, an "ensemble search" procedure may be employed. The simplest such procedure is an exhaustive search in which ω_ℓ is varied over a set of uniformly spaced "candidate frequencies" given by $\omega_c[i] = 2i\pi/M$ for $0 \le i \le M/2$. For each $\omega_c[i]$, the corresponding value of E_ℓ is determined by solving the normal equations, and ω_ℓ is chosen as that value of $\omega_c[i]$ yielding the minimum error. Having determined the parameters of the ℓth component, the successive approximation and error sequences are updated by (6), and the process is repeated for the next component.

Several points are important to note concerning this analysis procedure. First, recognizing that $e_\ell[n]$ may be viewed as a vector in the real Hilbert space $\boldsymbol{R}^{(2N_a+1)}$ and using results from vector space theory [17], it can be shown that if $\sigma[n]>0$ and $M>2N_a$, then the successive error norm E_ℓ will decrease monotonically with the addition of each new approximating component. This result provides an important guarantee of performance for analysis-by-synthesis by ensuring (under mild conditions) that synthetic speech must become more accurate given increasing numbers of components. As a practical matter, the value of M used in analysis-by-synthesis will be considerably larger than $2N_a$ to provide a fine grid of candidate frequencies. Of course, larger values of M imply more analysis computation, so the choice of M is a tradeoff. In order to provide a level of accuracy that is independent of the analysis frame length [17]–[19], M should be proportional to N_a, i.e., $M = \nu_a N_a$, where ν_a is typically in the range from three to eight, depending on the balance of quality requirements versus computational resources.

The analysis-by-synthesis algorithm just described was first proposed in [24] for constant-frequency sinusoidal components. Marques and Almeida [25] previously proposed a similar analysis algorithm that uses a generalized polynomial phase for sinusoidal components to boost accuracy, uses peak-picking analysis to determine component frequencies, and operates without a time-varying gain signal. We found this technique to be very effective for fine-grained speech analysis, but also found that the extra complexity required to analyze components with higher order polynomial phase was not justified by significant improvements in perceived speech quality for analysis/synthesis applications using moderate frame lengths.

Fig. 1 shows a functional block diagram of the analysis procedure described above, illustrating the "closed-loop" successive approximation structure reminiscent of linear predictive analysis-by-synthesis coders [22]. From (7) and Fig. 1, it is clear that by minimizing E_ℓ in terms of the parameters of the ℓth component, we are in fact approximating the residual error left after approximating the segment of $s[n]$ by the previous $\ell-1$ components.

In the preceding discussion, the recursive nature of analysis-by-synthesis implies that as many (or as few) components as desired may be calculated. The obvious problem that arises is how to determine the number of components required to model a given speech segment accurately. While a simple signal-to-noise ratio (SNR) based on E_ℓ provides a meaningful measure of signal approximation, it can lead to overapproximation of low-energy background noise during speaker inactivity. An energy-dependent threshold defined by analogy to the behavior of μ-law quantization is proposed in [17]; this approach deals effectively with overapproximation, and is expressed in terms of well-defined speech signal properties rather than arbitrary thresholds.

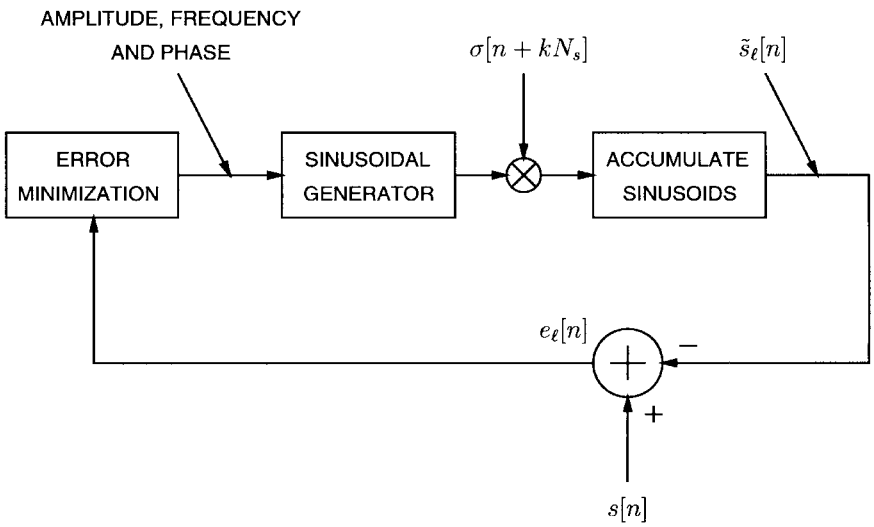

Fig. 1. Block diagram of analysis-by-synthesis procedure applied to overlap-add sinusoidal modeling.

1) Frequency-Domain Interpretation: Much of the computation required to determine overlap-add sinusoidal model parameters is in the form of inner products between sinusoids of various frequencies and between sinusoids and arbitrary discrete-time sequences [17]–[19]. As a result, much of the analysis-by-synthesis algorithm might be expected to be expressible in terms of frequency-domain operations and transforms. As we will see, this frequency-domain interpretation provides a great deal of useful information for the purposes of design and implementation of analysis-by-synthesis sinusoidal modeling.

The discrete-time Fourier transform (DTFT) of a sequence $x[n]$ is defined by

$$X(e^{j\omega}) \triangleq \sum_{n=-\infty}^{\infty} x[n]e^{-j\omega n}. \tag{10}$$

When $x[n]$ is a real-valued sequence, the following identities hold:

$$\sum_{n=-\infty}^{\infty} x[n]\cos\omega n = \Re e\{X(e^{j\omega})\}$$
$$\sum_{n=-\infty}^{\infty} x[n]\sin\omega n = -\Im m\{X(e^{j\omega})\}. \tag{11}$$

The DTFT's of the sequences $w_a[n]e_\ell[n]\sigma[n+kN_s], w_a[n]s[n+kN_s]\sigma[n+kN_s], w_a[n]\tilde{s}_\ell[n]\sigma[n+kN_s], w_a[n]\hat{s}_\ell[n]\sigma[n+kN_s]$, and $w_a[n]\sigma^2[n+kN_s]$ will be referred to as $EG_\ell(e^{j\omega}), SG(e^{j\omega}), \tilde{SG}_\ell(e^{j\omega})$ ("S-tilde"), $\hat{SG}_\ell(e^{j\omega})$ ("S-hat"), and $GG(e^{j\omega})$, respectively.

Equation (6) provides a recursive expression for the successive approximation sequence $\tilde{s}_\ell[n]$ and error sequence $e_\ell[n]$. Substituting the first relation of (6) into the formula for the DTFT $\tilde{SG}_\ell(e^{j\omega})$ and making use of (5) yields

$$\begin{aligned}\tilde{SG}_\ell(e^{j\omega}) &= \sum_{n=-N_a}^{N_a} w_a[n](\tilde{s}_{\ell-1}[n]+\hat{s}_\ell[n]) \\ &\quad\times\sigma[n+kN_s]e^{-j\omega n} \\ &= \tilde{SG}_{\ell-1}(e^{j\omega}) + \hat{SG}_\ell(e^{j\omega}).\end{aligned} \tag{12}$$

Likewise, $EG_\ell(e^{j\omega})$ may be expressed as

$$EG_\ell(e^{j\omega}) = EG_{\ell-1}(e^{j\omega}) - \hat{SG}_\ell(e^{j\omega}). \tag{13}$$

These relations imply a direct duality between recursively updating the successive approximation and error sequences and updating spectra associated with these sequences. According to the assumed initial conditions, $EG_0(e^{j\omega}) = SG(e^{j\omega})$ and $\tilde{SG}_0(e^{j\omega}) = 0$.

Likewise, substituting the definition of $\hat{s}_\ell[n]$ given by (5) into the expression for the "component spectrum" $\hat{SG}_\ell(e^{j\omega})$, it is easily shown that

$$\hat{SG}_\ell(e^{j\omega}) = \alpha_\ell GG(e^{j(\omega-\omega_\ell)}) + \alpha_\ell^* GG(e^{j(\omega+\omega_\ell)}) \tag{14}$$

where $\alpha_\ell = \frac{1}{2}A_\ell e^{j\phi_\ell}$. In other words, the component spectrum $\hat{SG}(e^{j\omega})$ is simply the conjugate sum of two identical spectra $GG(e^{j\omega})$ shifted left and right by ω_ℓ. When combined with (12), we see that the "successive approximation spectrum" $\tilde{SG}_\ell(e^{j\omega})$ is simply a weighted sum of shifted versions of $GG(e^{j\omega})$.

The projection theorem states that $e_\ell[n]$ must be orthogonal to $\hat{s}_\ell[n]$. Since $\hat{s}_\ell[n]$ is a linear combination of the sequences $\sigma[n+kN_s]\cos\omega_\ell n$ and $\sigma[n+kN_s]\sin\omega_\ell n$, this orthogonality condition may be expressed as

$$\sum_{n=-N_a}^{N_a} w_a[n]e_\ell[n]\sigma[n+kN_s]\cos\,\omega_\ell n = 0$$
$$\sum_{n=-N_a}^{N_a} w_a[n]e_\ell[n]\sigma[n+kN_s]\sin\,\omega_\ell n = 0.$$

Making use of (11), these orthogonality conditions are equivalent to the requirement that $EG_\ell(e^{j\omega_\ell}) = 0$; that is, under optimal conditions the "successive error spectrum" obtained by subtracting the component spectrum from the previous error spectrum as in (13) will have a spectral null at the component frequency ω_ℓ. Another interpretation follows by substituting (13):

$$EG_{\ell-1}(e^{j\omega_\ell}) = \hat{SG}_\ell(e^{j\omega_\ell}).$$

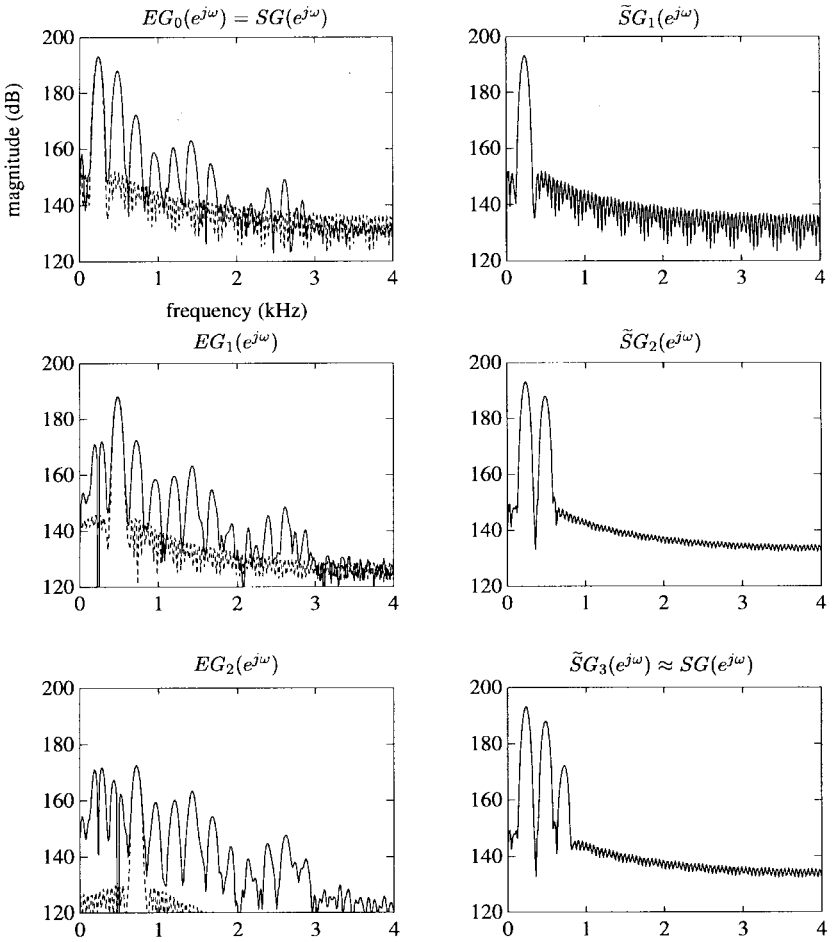

Fig. 2. Frequency-domain interpretation of analysis-by-synthesis.

This implies that, for a given component frequency ω_ℓ, the amplitude and phase parameters that minimize E_ℓ cause the component spectrum $\hat{S}G_\ell(e^{j\omega})$ to match the previous error spectrum $EG_{\ell-1}(e^{j\omega})$ at ω_ℓ. Fig. 2 shows an example of the operation of analysis-by-synthesis from a frequency-domain standpoint, using a 20 ms voiced speech segment. The left-hand plots in this figure show error spectra $EG_\ell(e^{j\omega})$, with optimal component spectra $\hat{S}G_\ell(e^{j\omega})$ dashed. The right-hand plots show approximation spectra $\tilde{S}G_\ell(e^{j\omega})$.

Fig. 2 illustrates two important (and related) features of analysis-by-synthesis. One difficulty noted in using windows to perform spectral analysis is that sidelobes in the window spectrum tend to cause interference or *cross-talk* between components, resulting in inaccurate estimates of component parameters [26]. Examining Fig. 2, analysis-by-synthesis is seen to remove the spectral effects of each component as it is estimated, counteracting cross-talk by "uncovering" the spectra of smaller components, thus providing more accurate synthetic speech.

This figure also illustrates that the frequency-domain behavior of analysis-by-synthesis is dependent on the analysis window spectrum $W_a(e^{j\omega})$. Using vector space principles, it has been demonstrated that the accuracy of analysis-by-synthesis deteriorates when signal components are not well resolved [17]. In the present context, this implies that if signal components are separated in frequency by less than half the mainlobe bandwidth of $W_a(e^{j\omega})$, the performance of analysis-by-synthesis will be poor.

In order to keep cross-correlation terms relatively small in cases of interest, it is important to choose an analysis window $w_a[n]$ such that $W_a(e^{j\omega})$ has a narrow mainlobe and sidelobes with relatively small magnitude, and to choose N_a such that the smallest differential frequency expected between any two signal components does not fall in the mainlobe of $W_a(e^{j\omega})$. Two good choices in terms of mainlobe width and sidelobe behavior are the Hamming and Kaiser $(\alpha = 6)$ windows; both have mainlobe widths of approximately $4\pi/N_a$ and relative sidelobe magnitudes less than -40 dB.

Addressing the problem of frequency resolution requires some knowledge of the signal being analyzed. When dealing with voiced speech, which possesses a quasiharmonic structure, it is expected that the minimum frequency between any two signal components is the fundamental frequency ω_o. As suggested by McAulay and Quatieri [12], the ABS/OLA system adapts the value of N_a such that the window length $2N_a + 1$ is 2.5 average pitch periods long but no shorter than 20 ms.

C. Perceptual Considerations

The analysis-by-synthesis procedure defined in this section makes use of a mean-square error criterion to determine model parameters. As we have seen, least-squares sinusoidal

model analysis leads to tractable, closed-form solutions, and allows analysis to be cast in terms of familiar frequency-domain concepts. For audio signals, however, mean-square error measures do not correlate well with subjective measures of fidelity as perceived by human listeners [27]; as a result, while analysis-by-synthesis as described can come very close to achieving the highest SNR possible for a given number of component sinusoids, this may not imply that perceived quality is as high as possible. The reasons for this are seen by considering the analysis-by-synthesis example of Fig. 2. After the first component spectrum $\hat{S}G_1(e^{j\omega})$ is removed to produce the error spectrum $EG_1(e^{j\omega})$, there is a significant amount of spectral energy within a mainlobe bandwidth of the first component frequency, due to slight nonstationarities and background noise. As a result, analysis-by-synthesis will tend to "cluster" low-amplitude components near the first component frequency, rather than choosing more perceptually important low-energy components at higher frequencies.

Given the observation that low-amplitude sinusoids cluster around high-amplitude sinusoids mainly within the mainlobe bandwidth of $W_a(e^{j\omega})$, a simple but effective heuristic is suggested: Once a component with frequency ω_ℓ is determined, frequencies in the range

$$\omega_\ell - \gamma_b \frac{B_{\rm ml}}{2} \leq \omega \leq \omega_\ell + \gamma_b \frac{B_{\rm ml}}{2}$$

where $B_{\rm ml}$ is the mainlobe bandwidth, may be removed from the ensemble search thereafter. Since analysis-by-synthesis tends to choose components in order of decreasing energy [17], this approach eliminates the problem of clustering with minimal impact on the overall analysis. This "frequency blanking" method, which requires no computational overhead, very effectively reduces clustering and achieves perceptual results similar to that of using a perceptual weighting filter [17], [19]. Experiments indicate that $\gamma_b = .75$ yields the best perceptual results, but this value is not critical.

III. Speech Modification Using the ABS/OLA System

Having described how an overlap-add sinusoidal speech model and an analysis-by-synthesis procedure may be combined to form an accurate analysis/synthesis system, what remains is to apply this system to the problem of speech modification. This section discusses the issues involved in applying the ABS/OLA system to time-, frequency-, and pitch-scale modification of speech. Particular emphasis is placed on a refined quasiharmonic overlap-add sinusoidal model formulation that is shape invariant in the presence of modifications, and on a pitch-scale modification algorithm that addresses the problems of bandwidth compression and noise migration.

A. Quasiharmonic Modeling

Before proceeding, it is important to introduce an overlap-add sinusoidal model formulation that will be useful for speech modification. While the sinusoidal model described in Section II is capable of producing approximations to speech signals that are perceptually identical to the originals, for the purpose of speech modification a synthetic contribution should reflect pitch information embedded in the signal. To this end, the synthetic contribution $\tilde{s}^k[n]$ given in (2) may be rewritten in "quasiharmonic" form as

$$\begin{aligned}\tilde{s}^k[n] &= \sum_{\ell=0}^{J[k]} A_\ell^k \cos(\omega_\ell^k n + \phi_\ell^k) \\ &= \sum_{\ell=0}^{J[k]} A_\ell^k \cos((\ell\omega_o^k + \Delta_\ell^k)n + \phi_\ell^k) \end{aligned} \tag{15}$$

where $J[k]$ is now the greatest integer such that $J[k]\omega_o^k \leq \pi$. Note that only one component is associated with each harmonic number ℓ, but that each component frequency has an arbitrary value expressed in terms of its differential frequency Δ_ℓ^k. Therefore, although pitch information is used in the formulation, the quasiharmonic sinusoidal model does not result in a "pitch-driven" analysis/synthesis system.

With this model formulation, it is necessary to calculate the fundamental frequency ω_o^k associated with a synthetic contribution as well as the amplitudes, frequencies, and phases of $\tilde{s}^k[n]$ in each analysis frame. McAulay and Quatieri have proposed a pitch estimation algorithm that operates on sinusoidal model parameters by evaluating a range of candidate fundamental frequencies in terms of a mean-square error criterion [28]. A similar but less complex algorithm proposed in [17], [19] constrains the search space of candidate fundamental frequencies based on knowledge of component frequencies gained from analysis-by-synthesis. Given a rough fundamental frequency estimate, it is possible to simultaneously arrange a subset of the model parameters from analysis-by-synthesis in the quasiharmonic form of (15) and recursively refine the fundamental frequency estimate. It should be noted that fundamental frequency estimation in the ABS/OLA system is, in general, independent of analysis-by-synthesis. In addition, the quasiharmonic representation described is not particularly sensitive to the pitch estimation algorithm used to determine ω_o^k.

B. Time- and Frequency-Scale Modification

As discussed in Section I, sinusoidal modeling is well-suited to the task of independently accessing and modifying information corresponding to the perceptual properties of speech. The sinusoidal transformation system of McAulay and Quatieri synthesizes speech using sinusoids whose amplitude and frequency vary in a piecewise continuous manner over time, providing the means to modify speech by manipulation of amplitude and frequency tracks. However, since the ABS/OLA synthesis model assumes no explicit relationship between frames, such continuous parameter functions are not available.

Speech modifications in the ABS/OLA system must instead be accomplished by modifying individual synthetic contributions $\tilde{s}^k[n]$ in (1) to achieve desired changes in time and frequency scale. In addition, the modification system must exhibit both *intraframe shape invariance*, which is the ability to maintain the underlying shape of the original speech waveform in the presence of modification [14], [15], and *interframe coherence*, defined as the ability to maintain correct temporal

phase evolution when modified contributions are summed together [14]. Accomplishing these goals requires both a generalized overlap-add model formulation and algorithms designed to insure coherence in the modified signal from frame to frame. These features of the ABS/OLA system are discussed in the following subsections.

1) A Refined Overlap-Add Modification Model: As previously noted, an approach to speech modification using an overlap-add model formulation (namely the DSTFT) has been reported by Portnoff [1]. Initially, a similar strategy to perform time- and frequency-scale modification was devised for the ABS/OLA system using the quasiharmonic overlap-add sinusoidal model [17]–[19]. The results of this modification scheme were not encouraging, particularly for time or frequency scale factors greater than one. Although waveform shape was generally preserved to a greater degree than in modification using the DSTFT, synthetic speech waveforms were found to break down quickly when extrapolated beyond the original analysis interval.

To maintain the basic overlap-add structure while improving the shape-invariance of modified speech produced by the ABS/OLA system, it was therefore necessary to develop a modification model formulation similar to that used for shape-invariant modification in the STS [15], but more specifically suited to the overlap-add model. Given a spectral envelope estimate $H^k(e^{j\omega})$ of the speech signal at $n = kN_s$, we associate each synthetic contribution from the quasiharmonic representation of (15) with an *excitation contribution* of the form

$$\tilde{e}^k[n] = \Re e\left\{\sum_{\ell=0}^{J[k]} b_\ell^k e^{j\Delta_\ell^k n} e^{j(\ell\omega_o^k n + \theta_\ell^k)}\right\}. \tag{16}$$

The fixed amplitude and phase parameters of $\tilde{e}^k[n]$ are given by

$$\begin{aligned} b_\ell^k &= A_\ell^k / |H^k(e^{j\omega_\ell^k})| \\ \theta_\ell^k &= \phi_\ell^k - \angle H^k(e^{j\omega_\ell^k}). \end{aligned} \tag{17}$$

These operations act to remove the effects of the vocal tract in the sense of frequency-domain deconvolution [29].

In this model formulation, the magnitude and phase changes of $H^k(e^{j\omega})$ from frame to frame, coupled with the amplitude modulation imparted by the envelope sequence $\sigma[n]$ and synthesis windows $\{w_s[n - kN_s]\}$, may be assumed to represent slowly varying amplitude and phase modulations due to vocal tract variations in the speech signal. Therefore, in order to change the articulation rate of analyzed speech, the time scale of these modulation terms should be altered during time-scale modification and left unchanged during frequency-scale modification, with no time shift imparted in either case. Furthermore, pitch information in this model is assumed to be represented entirely by the fundamental frequency ω_o^k, thus scaling the harmonic terms of (16), produces a desired change in frequency scale.

Closer examination of (16) reveals complex modulation terms imparted to each component by the differential frequencies $\{\Delta_\ell^k\}$. Assuming that the differential frequency terms are relatively small compared to the fundamental frequency ω_o^k [13], we may interpret these complex modulation terms as contributions to the evolution of the vocal tract magnitude and phase over time, rather than assuming them to be coupled to the harmonic frequencies. Based on the described role of amplitude/phase modulation in modification, the time scale of these factors should thus be altered during time-scale modification and left unchanged during frequency-scale modification, with no time shift imparted in either case.

Quantitatively, to time-scale modify synthesis frame k by a factor of ρ_k and frequency-scale by β_k, the above analysis implies replacing the term $\Delta_\ell^k n$ with $\Delta_\ell^k n/\rho_k$ in (15). The harmonic frequencies of each synthetic contribution are then scaled to produce a fundamental frequency change, and a global time shift is imparted to the harmonic components in order to preserve interframe phase coherence in the modified speech signal; this corresponds to replacing the term $\ell\omega_o^k n$ with $\ell\beta_k\omega_o^k(n + \delta^k)$. These observations may be combined to construct a synthesis equation that produces a modified synthesis frame [17]–[19]

$$\begin{aligned} \hat{s}[n + N_k] = \sigma\left[\frac{n}{\rho_k} + kN_s\right]\Bigg(& w_s\left[\frac{n}{\rho_k}\right]\tilde{s}^k_{\rho_k,\beta_k}[n] \\ &+ w_s\left[\frac{n}{\rho_k} - N_s\right]\tilde{s}^{k+1}_{\rho_k,\beta_{k+1}}[n - \rho_k N_s]\Bigg) \end{aligned} \tag{18}$$

for $0 \le n < \rho_k N_s$, where $N_k = N_s \Sigma_{i=0}^{k-1} \rho_i$ is the starting point of the modified synthesis frame, and where modified synthetic contributions are expressed as

$$\begin{aligned} \tilde{s}^k_{\rho_k,\beta_k}[n] &= \sum_{\ell=0}^{J[k]} A_\ell^k \cos\left(j\beta_k\omega_o^k(n + \delta^k) + \frac{\Delta_\ell^k n}{\rho_k} + \phi_\ell^k\right) \\ \tilde{s}^{k+1}_{\rho_k,\beta_{k+1}}[n] &= \sum_{\ell=0}^{J[k+1]} A_\ell^{k+1} \cos\Bigg(j\beta_{k+1}\omega_o^{k+1}(n + \delta^{k+1}) \\ &\qquad + \frac{\Delta_\ell^{k+1} n}{\rho_k} + \phi_\ell^{k+1}\Bigg). \end{aligned} \tag{19}$$

It may be noted that these synthesis equations impart time-scale modification to each synthesis frame, but frequency-scale modification to each synthetic contribution. This is done to facilitate time-varying modification using the overlap-add model; time-scale modification on each synthesis frame ensures a consistent overlap of each contribution, while frequency-scale modification of each contribution avoids phase discontinuities in the presence of rapid frequency modifications [17]. It should also be noted that these synthesis equations may, in general, require the evaluation of discrete-time sequences at noninteger indices. In the case of the synthesis window $w_s[n]$ and synthetic sequence $\tilde{s}^k_{\rho_k,\beta_k}[n]$ this is not a problem, since these sequences are functionally defined at any point. Furthermore, since the time-varying gain sequence $\sigma[n]$ is a low bandwidth sequence, linear interpolation is effective to produce required sequence values.

While this approach to modification bears similarity to overlap-add modification using the DSTFT and to shape-invariant modification in the STS, there is a critical difference

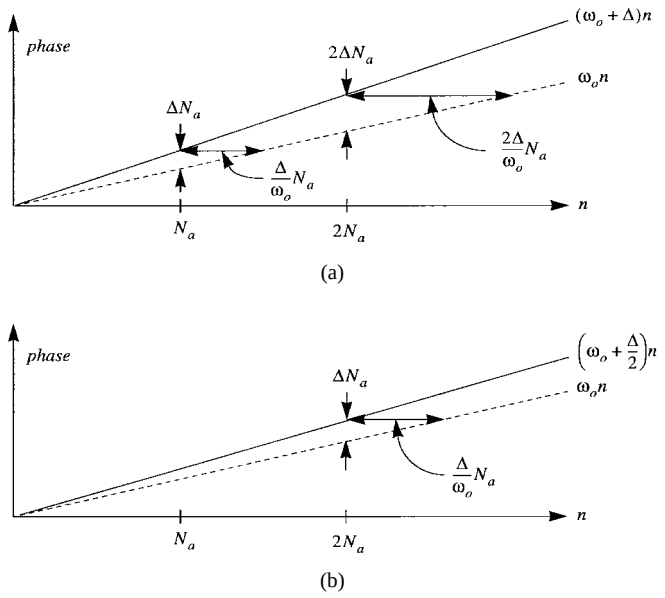

Fig. 3. Intraframe phase coherence for single component (a) without differential frequency scaling and (b) with differential frequency scaling.

as well: The component frequencies of each synthetic contribution are altered according to the relation

$$\hat{\omega} = \ell\beta\omega_o + \Delta_\ell/\rho. \tag{20}$$

This implies that as the time scale factor ρ is increased (corresponding to time-scale expansion), the component frequencies of $\tilde{s}^k_{\rho_k,\beta_k}[n]$ are "pulled in" toward the harmonic frequencies, and in the limit become harmonically related.

To understand the significance of this frequency modification, consider how the differential frequencies $\{\Delta^k_\ell\}$ affect the intraframe phase relation of the components of $\tilde{s}^k[n]$. Fig. 3 demonstrates the phase behavior of a component with frequency $\omega_o + \Delta$; Fig. 3(a) shows the component's linear phase and the "harmonic reference" line with slope ω_o. Clearly, as a function of n, the component phase leads that of the harmonic reference by Δn radians, and leads in time by $\Delta n/\omega_o$ samples. At the analysis frame boundary $(n = N_a)$, these leads have the values shown. As illustrated in Fig. 3(a), the effect of the differential frequency term is a linearly increasing phase and time misalignment with respect to the harmonic reference. Since each component sinusoid exhibits similar behavior, and since the differential frequencies are unrelated in general, these increasing leads cause the components of $\tilde{s}^k[n]$ to lose phase lock as n becomes large. As a result, the waveform shape of $\tilde{s}^k[n]$ becomes distorted as the time index is extrapolated beyond analysis frame boundaries.

Fig. 3(b) shows how the frequency modification of (20) affects the same component's phase behavior for time-scale modification by a factor of two. As before, the phase and time leads of this component are proportional to the modified differential frequency, but since this differential is half of the unmodified differential, the resulting phase and time leads increase at half the rate of the unmodified component, and at $n = 2N_a$ are the same as the unmodified component at $n = N_a$. The effect of the frequency modification of (20), therefore, is both to constrain the modified contribution $\tilde{s}^k_{\rho_k,\beta_k}[n]$ to have the same underlying waveform shape as $\tilde{s}^k[n]$ at the modified frame boundary $n = \rho_k N_s$ and to control the rate of phase deviation to account for variable synthesis frame lengths. As a result, manipulating the differential frequencies preserves the original synthetic contribution's underlying waveform shape over the modified synthesis frame [17]–[19].

2) Interframe Coherence: The time shifts δ^k and δ^{k+1} of the modified synthetic contributions may be specified in terms of constraints designed to preserve the coherence of synthetic contributions from frame to frame. In both the STS and the ABS/OLA system, interframe coherence is quantified in terms of the "pitch onset time" excitation model developed by McAulay and Quatieri [30]. In this model, the excitation contribution of (16) may be rewritten as

$$\tilde{e}[n] = \sum_{\ell=0}^{J} b_\ell \cos(\omega_\ell(n - \tau_p) + \psi_\ell(\tau_p)) \tag{21}$$

where frame notation is again suppressed, and where $\psi_\ell(\tau_p) = \theta_\ell + \omega_\ell\tau_p$.

According to the source/filter model for voiced speech, $\tilde{e}[n]$ is expected to correspond approximately to a sequence of impulses separated by a pitch period of $2\pi/\omega_o$ samples. Equation (21) parameterizes this behavior in terms of the "pitch onset time" τ_p at which a pitch pulse occurs relative to $n = 0$. Under this assumption, at $n = \tau_p$ the components of $\tilde{e}[n]$ will add coherently, implying that the time shifted "residual" phase parameters $\{\psi_\ell(\tau_p)\}$ will all be close to either zero or π or *maximally coherent*. Based on this observation, McAulay and Quatieri have proposed a technique to estimate τ_p [30] by maximizing a *pitch onset likelihood function*. An extension of this technique, reported in [17] and [19], approaches this problem as one of matching the speech signal with the signal produced by driving the vocal tract filter with an ideal pulse train offset by τ_p samples.

As previously mentioned, the ideal glottal excitation waveform for voiced speech is a variable-frequency pulse train; to produce such a structured waveform using an overlap-add model, the pulse locations of the synthetic contributions given by (2) must be highly correlated from one frame to the next. Therefore, in the presence of modifications this correlation must be maintained if the resulting modified speech is to be free from artifacts. To accomplish this, the time shifts δ^k and δ^{k+1} in (19) may be determined such that the underlying excitation signal exhibits the desired level of correlation when modified [14], [15].

In the ABS/OLA system, this is accomplished using an extension of the onset time modification algorithm of the STS; the result of this coherence preservation algorithm is a closed-form, recursive relation for δ^{k+1} given δ^k [17], [19]. These time shifts serve to produce a modified speech waveform that exhibits no unexpected pitch period fluctuations that would be detected as artifacts. As in the STS [15], this coherence algorithm can be shown to be insensitive to pitch estimation errors such as multiplication or division by an integer factor [17]. Since pitch estimators will always make such gross errors, this result is extremely important in terms of modified speech quality.

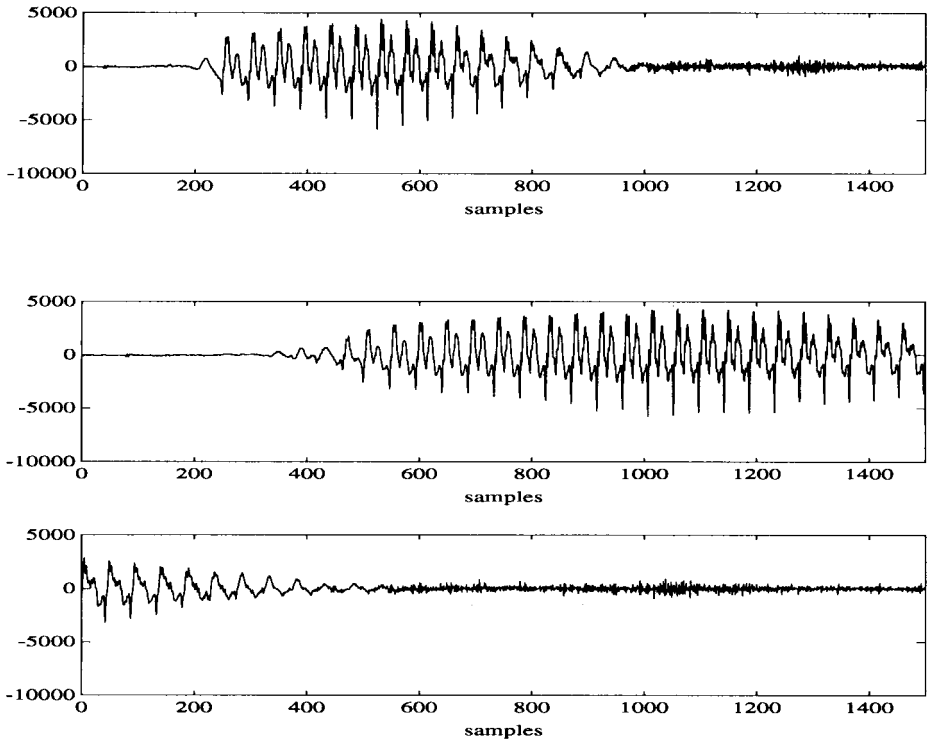

Fig. 4. Waveform plots of ABS/OLA time-scale modification. (a) Original waveform of word "rust." (b) Waveform time-scale modified by factor $\rho = 2$.

Fig. 4 shows an example of time-scale expansion using the modification scheme just described. Fig. 4(a) is a plot of the word "rust" spoken by a female talker; Fig. 4(b) illustrates the synthetic waveform produced by the ABS/OLA system for time-scale expansion by a factor of two. The utterance was modeled using a synthesis frame length (analysis frame interval) of 10 ms and a 20 ms Hamming analysis window. This example clearly demonstrates the ability of the ABS/OLA system both to model significant features of the speech signal accurately and to produce a modified speech signal whose waveform characteristics match the original.

Fig. 4 also demonstrates the ability of the ABS/OLA system to capture and modify nonvoiced speech events (including plosives and stops) as well, despite the fact that the modification algorithm was designed with voiced speech in mind. This is not terribly surprising for several reasons. First, sinusoidal models are known to represent unvoiced speech well, provided a sufficient number of components with random frequency and phase are used [12]. Second, although quasiharmonic structure is imposed on contributions used to synthesize unvoiced speech, both the phase and fundamental frequency of adjacent contributions are uncorrelated, implying that synthesized unvoiced speech maintains its random characteristics. Finally, the algorithm used in the ABS/OLA system to preserve interframe coherence is designed to maintain whatever level of correlation exists between synthetic contributions used in the synthesis model. Since the contributions corresponding to unvoiced speech events are uncorrelated, modification in the ABS/OLA system does not impose structure on unvoiced speech in the presence of modifications.

As in the STS, an important caveat concerning unvoiced speech is the need to maintain reasonably short synthesis frame lengths to ensure randomness in modified unvoiced speech; experiments indicate that modified synthesis frames 20 ms or shorter are sufficient to maintain natural synthetic speech quality. Furthermore, it should be noted that extreme time-scale expansion of fricative sounds results in tonal artifacts, due to the fact that component frequencies in the ABS/OLA system become harmonic as $\rho \rightarrow \infty$.

Fig. 5 illustrates the behavior of ABS/OLA time-scaling in the frequency domain using wideband spectrograms of the sentence "Line up at the screen door" spoken by a low-pitched male talker. Fig. 5(a) shows the spectrogram of the original sentence, while Fig. 5(b) shows the spectrogram of the utterance time-scaled by a factor of two. As in the waveform plot, this figure demonstrates that only the desired change in articulation rate takes place in ABS/OLA processing, and that both the pitch and formant structure of the original speech remain intact.

Time-scale modification using the ABS/OLA system has been applied to speech from a variety of English-speaking male and female talkers over a range of background noise conditions using both conversational and "read speech" data bases. The ABS/OLA system has performed well in these experiments, producing high-quality modified speech free from noticeable artifacts. In one experiment, air traffic control transmissions intercepted from a noisy, fading communication channel were slowed to half-speed using the ABS/OLA system, with a dramatic improvement in intelligibility and without objectionable artifacts. Fixed frequency-scale modification has

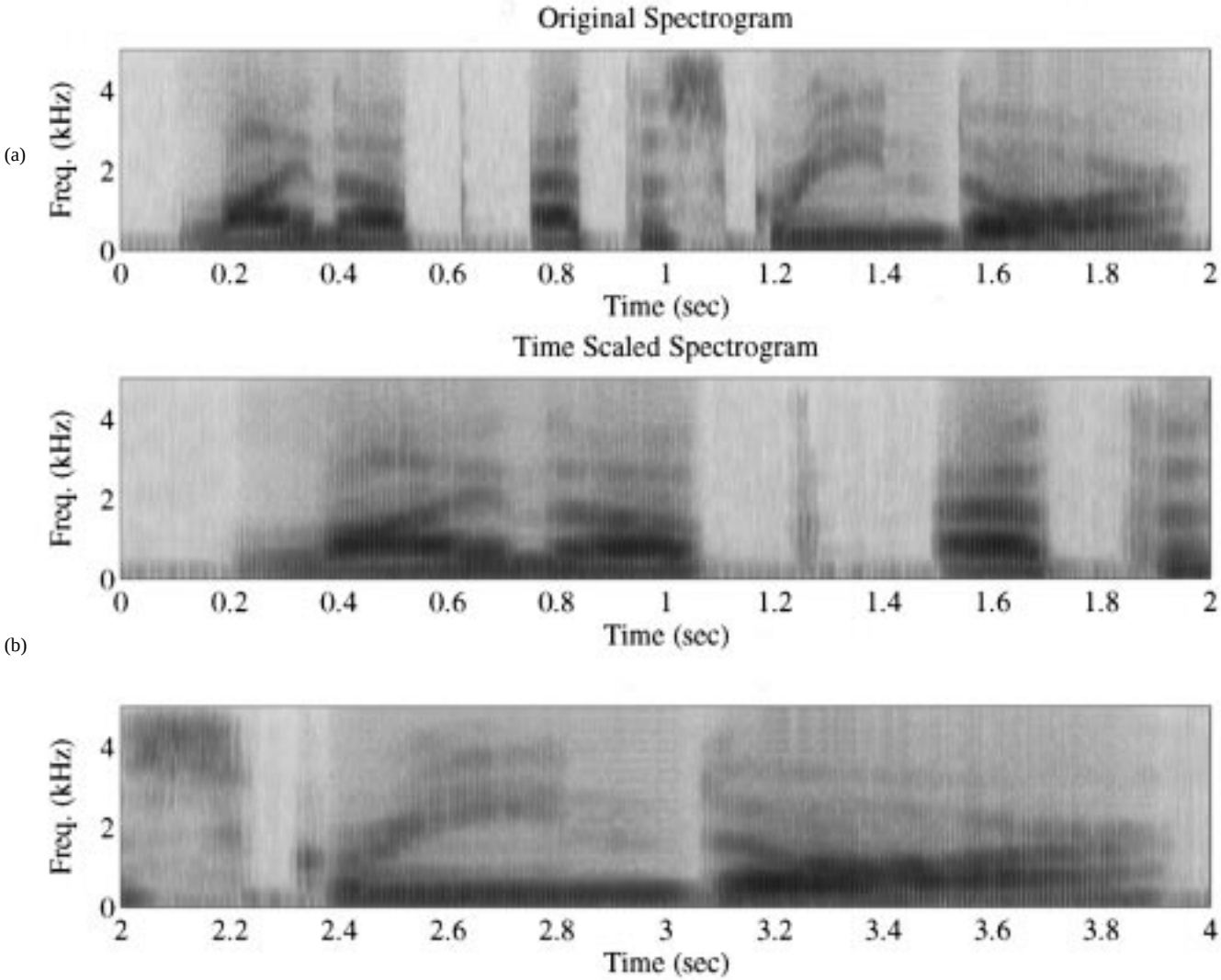

Fig. 5. Spectrographic plots of ABS/OLA time-scale modification. (a) Wideband spectrogram of sentence, *Line up at the screen door*. (b) Spectrogram of sentence time-scaled by factor of two.

also been successfully implemented with similar robustness over speakers and speaking environments; a detailed discussion is presented later.

Implicit in the described modification algorithm is the ability to perform joint time-varying time- and frequency-scale modification. Various experiments have been performed using the ABS/OLA system to impart joint time-frequency modifications and time-varying time and frequency scaling, with results similar to those for fixed time-scale modification. In one experiment, the ABS/OLA system was used very effectively to dynamically time warp utterances for temporal matching.

Experiments indicate that while modification constraints must be placed on unvoiced speech events, the limitations of voiced speech modification are much less strict. Since the refined modification model supports arbitrarily long synthesis frames, the only restriction for time scaling voiced speech is to ensure that the analysis frame interval is short enough to model rapidly spoken phonemes. Intervals on the order of 10 ms are generally sufficient for time-scale compression or expansion up to a factor of two. Unlike the STS, however, larger time scales do not require proportionally shorter frame intervals; given an interval of 5 ms, voiced speech time scaling up to a factor of eight has been successfully attempted using the ABS/OLA system, without noticeable artifacts.

C. Pitch-Scale Modification

Frequency-scale modification can successfully change the fundamental frequency of analyzed speech without changing its time scale or introducing artifacts. However, when the goal of modification is natural-sounding pitch alteration, frequency-scale modification has significant disadvantages. For instance, when scale factors less than one are used, the modified component frequencies span a proportionally smaller range. This *bandwidth compression* results in a loss of high-frequency energy and imparts a "muffled" quality to the modified speech. Worse yet, the component amplitudes $\{A_\ell^k\}$ are unaltered by frequency-scale modification, resulting in an altered formant structure that degrades both message content and speaker identity. For these reasons it is important to consider an approach to pitch-scale modification that alters the fundamental frequency of analyzed speech while preserving its original formant structure.

An approach to pitch-scale modification has been proposed for the sinusoidal transformation system in [13] and [15]. As discussed in the preceding section, the pitch onset time model represents the glottal excitation waveform using a sinusoidal model whose parameters are determined by the relations of (17). Since the excitation waveform is a spectrally flat, quasiperiodic signal with pitch-pulse dispersion parameterized by residual phase parameters $\{\psi_\ell(\tau_p)\}$, the approach suggested by Quatieri and McAulay is to add the residual phase parameters to those of a spectral envelope estimate to produce a "mixed-phase" estimate at the measured component frequencies [29]. Complex interpolation of the spectral envelope is used to derive a new set of residual phase parameters at component frequencies modified by β_k. The excitation magnitudes are then multiplied by the measured spectral envelope magnitude at the modified frequencies, and the resulting parameter set used to generate pitch-modified speech.

While this approach succeeds in preserving the formant structure of modified speech, two problems are apparent. First, since the original set of component frequencies are scaled as in frequency-scale modification, the problem of high-frequency energy loss is not addressed, implying that pitch-lowered speech still sounds muffled. Second, noise added to components in low energy interformant regions may be highly amplified if pitch-scale modification moves those component frequencies near a formant. This "noise amplification" property can seriously affect the quality of pitch-modified speech. What follows is a description of an alternate approach to pitch-scale modification used in the ABS/OLA system that addresses these difficulties.

Consider a single synthetic contribution to the excitation sequence $e[n]$ given by (16). The objective in modifying the fundamental frequency of $\tilde{e}^k[n]$ without changing its spectral balance is to specify a set of parameters for a modified excitation contribution given by

$$\hat{e}^k_{\beta_k}[n] = \sum_{\ell=0}^{\hat{J}[k]} \hat{b}^k_\ell \cos(\beta_k \ell \omega^k_o (n+\delta^k) + \hat{\Delta}^k_\ell n + \hat{\theta}^k_\ell) \qquad (22)$$

such that the component frequencies of $\hat{e}^k_{\beta_k}[n]$ span the range $[0,\pi]$.

In the ABS/OLA system, this goal is achieved by interpolating the complex *phasor form* of excitation amplitude/phase pairs, expressed as $b_\ell e^{j\psi_\ell(\tau_p)}$. Given the quasiharmonic structure of the ABS/OLA modification model, it is reasonable to uniformly interpolate these phasor values at harmonic frequencies to produce a continuous excitation spectrum $\mathcal{E}_k(\omega)$. Given a pitch-scale factor β_k, the excitation spectrum is then resampled at modified harmonic frequencies $\beta_k \ell \omega^k_o$ in the range $[0,\pi]$ to generate phasor values for $\hat{e}^k_{\beta_k}[n]$. Note that residual phase parameters $\{\psi_\ell(\tau_p)\}$ are used in this "phasor interpolation" procedure. The primary reasons for this choice are the proximity of residual phase terms to zero or π, which causes phasor interpolation to be approximately linear in terms of excitation magnitudes, and their independence from pitch onset time relative to synthesis frame boundaries [17]–[19].

Quantitatively, the excitation spectrum is given by

$$\mathcal{E}(\omega) = \sum_{\ell=0}^{J} b_\ell e^{j\psi_\ell(\tau_p)} I(\omega - \ell\omega_o; \omega_o) \qquad (23)$$

where frame notation is again suppressed. Phasor interpolation yields modified excitation magnitudes given by $\hat{b}_\ell = |\mathcal{E}(\beta\ell\omega_o)|$ and modified residual phases given by $\hat{\psi}_\ell(\tau_p) = \angle\mathcal{E}(\beta\ell\omega_o)$.

To protect against noise amplification, a bandlimited interpolation function $I(\omega;\omega_o)$ is particularly useful. If $I(\omega;\omega_o)$ is bandlimited, the effect of any single noise-corrupted component of $\tilde{e}^k[n]$ on the modified excitation parameters is limited to the immediate neighborhood of that component's frequency. This greatly reduces the problem of noise amplification by ensuring that noise effects in one part of the speech spectrum do not migrate to another part of the spectrum upon modification. Stated another way, bandlimited interpolation ensures that random fluctuations in $\mathcal{E}(\omega)$ caused by noisy harmonics at the original fundamental frequency are not "picked up" in formant regions by harmonics at the modified fundamental frequency.

After phasor interpolation, modified excitation phases used in synthesis are derived by imparting a corrective time shift to the modified residual phase terms [17]–[19], i.e., $\hat{\theta}^k_\ell = \hat{\psi}^k_\ell(\tau^k_p) - \ell\omega^k_o\tau^k_p$. At this point, what remains is to specify appropriate differential frequency terms in the equation for $\hat{e}^k{}_{\beta_k}[n]$. Although this task is somewhat arbitrary, it is reasonable to expect that the differential frequency terms may be interpolated uniformly in a manner similar to phasor interpolation. This has the effect that modified differential frequencies follow the same trend in the frequency domain as the unmodified differentials, which is important both in preventing migration of noise effects and in modifying voiced fricatives and other speech events possessing a noisy structure in certain portions of the spectrum [21].

Given the amplitude, phase, and differential frequency parameters of the modified residual, the specification of a synthetic contribution to pitch-scale modified speech is completed by reintroducing the effects of the spectral envelope to the amplitude and phase parameters at the modified frequencies $\hat{\omega}^k_\ell = \beta\ell\omega^k_o + \hat{\Delta}^k_\ell$:

$$\hat{A}^k_\ell = \hat{b}^k_\ell |H^k(e^{j\hat{\omega}^k_\ell})|$$
$$\hat{\phi}^k_\ell = \hat{\theta}^k_\ell + \angle H^k(e^{j\hat{\omega}^k_\ell}).$$

The determined parameter set may now be used in the modification synthesis of (18) and (19), using the same time shift parameters δ^k and δ^{k+1} calculated for time- and frequency-scale modification.

Fig. 6 demonstrates the effects of frequency- and pitch-scale modification using the ABS/OLA system on a voiced speech waveform segment. Fig. 6(a) illustrates 500 samples of the phoneme /u/ as in the word "up." Fig. 6(b) shows the synthetic waveform segment produced by the ABS/OLA system for a frequency-scale factor $\beta = .75$. The synthetic segment is clearly reduced in pitch frequency, and appears to be an interpolated version of the original waveform due to the shape-invariance property of the refined modification model. In the

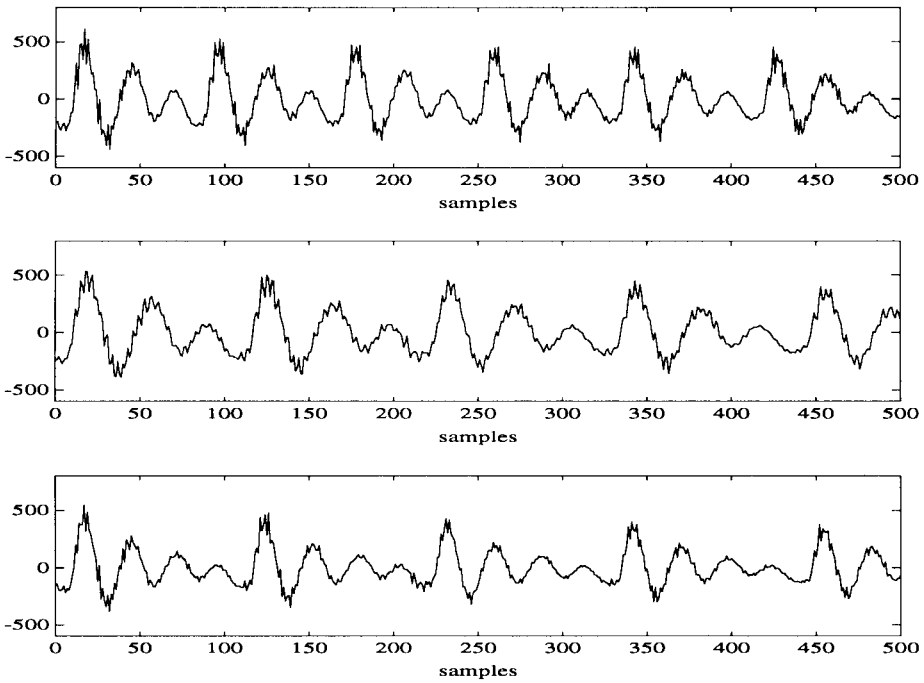

Fig. 6. Waveform plots of ABS/OLA frequency- and pitch-scale modification. (a) Original voiced speech waveform segment. (b) Frequency-scaled segment ($\beta = .75$); (c) Pitch-scaled segment.

context of pitch alteration, however, waveform interpolation reflects the undesirable spectral distortion and loss of high-frequency energy caused by frequency-scale modification.

By contrast, Fig. 6(c) shows the waveform segment produced by the ABS/OLA system for pitch-scale modification by the same scale factor. In this case, the waveform does not appear to be interpolated, but has a shape similar to the original with larger pitch period spacing. Furthermore, the modified waveform appears to have high frequency energy comparable to the original, demonstrating the ability of phasor interpolation to preserve the original speech bandwidth in the presence of pitch scaling.

Fig. 7 illustrates ABS/OLA frequency- and pitch-scale modification in the frequency domain: Fig. 7(a) is a narrowband spectrogram of the sentence "This road forks up ahead;" the harmonic component frequencies are indicated clearly by the dark horizontal bands. Fig. 7(b) shows the spectrogram of this sentence frequency scaled by the ABS/OLA system using $\beta = .75$. This spectrogram demonstrates both the desired change in fundamental frequency and time evolution of speech events similar to the original, but also clearly shows the undesirable compression of spectral information and loss of information at high frequencies.

Fig. 7(c) shows the spectrogram of this sentence after pitch-scale modification using the ABS/OLA system with the same scale factor as above. As in the case of frequency-scale modification, the closer spacing of harmonic components demonstrates the desired change in fundamental frequency. Unlike the frequency scaled speech, however, the spectrogram of pitch-scale modified speech has the same formant tracks as the original speech and exhibits no bandwidth compression. This behavior implies that the ABS/OLA system successfully alters the fundamental frequency of the analyzed utterance without degrading its intelligibility or significantly altering speaker identity.

As with time-scale modification, frequency- and pitch-scale modification using the ABS/OLA system have been extensively tested on a variety of speech utterances over a wide range of environments, with very satisfactory results. Experiments indicate that these modifications can be performed reliably using both clean speech and speech with considerable background and channel distortions as well; this is due largely to the ability of phasor interpolation to avoid noise amplification. Pitch-scale modification at moderate scale factors produces modified speech that sounds very much like the original speaker at the modified pitch. In addition, the ability of phasor interpolation to preserve speech bandwidth greatly improves the naturalness of pitch-modified speech. Time-varying pitch-scale modification has been used to impart a variety of pitch contours to analyzed speech, including the generation of monotone speech, without imparting objectionable artifacts.

While the ABS/OLA system produces desirable pitch alterations in analyzed speech, there are some limitations. As in the case of time-scale modification, unvoiced speech is not pitch-scale modified well. This is due both to the lack of true pitch information and to the inappropriateness of the pitch onset time model; the effect is pronounced in pitch-scale modification, since pitch changes can impart undesirable structure to modified unvoiced speech. As in the STS, suspending modification on the basis of voiced/unvoiced classification significantly reduces this distortion [15].

Another effect of ABS/OLA pitch modification is that synthetic speech tends to sound less like the original talker for large pitch-scale modifications. This is because phasor interpolation fails to accurately predict vocal tract changes in

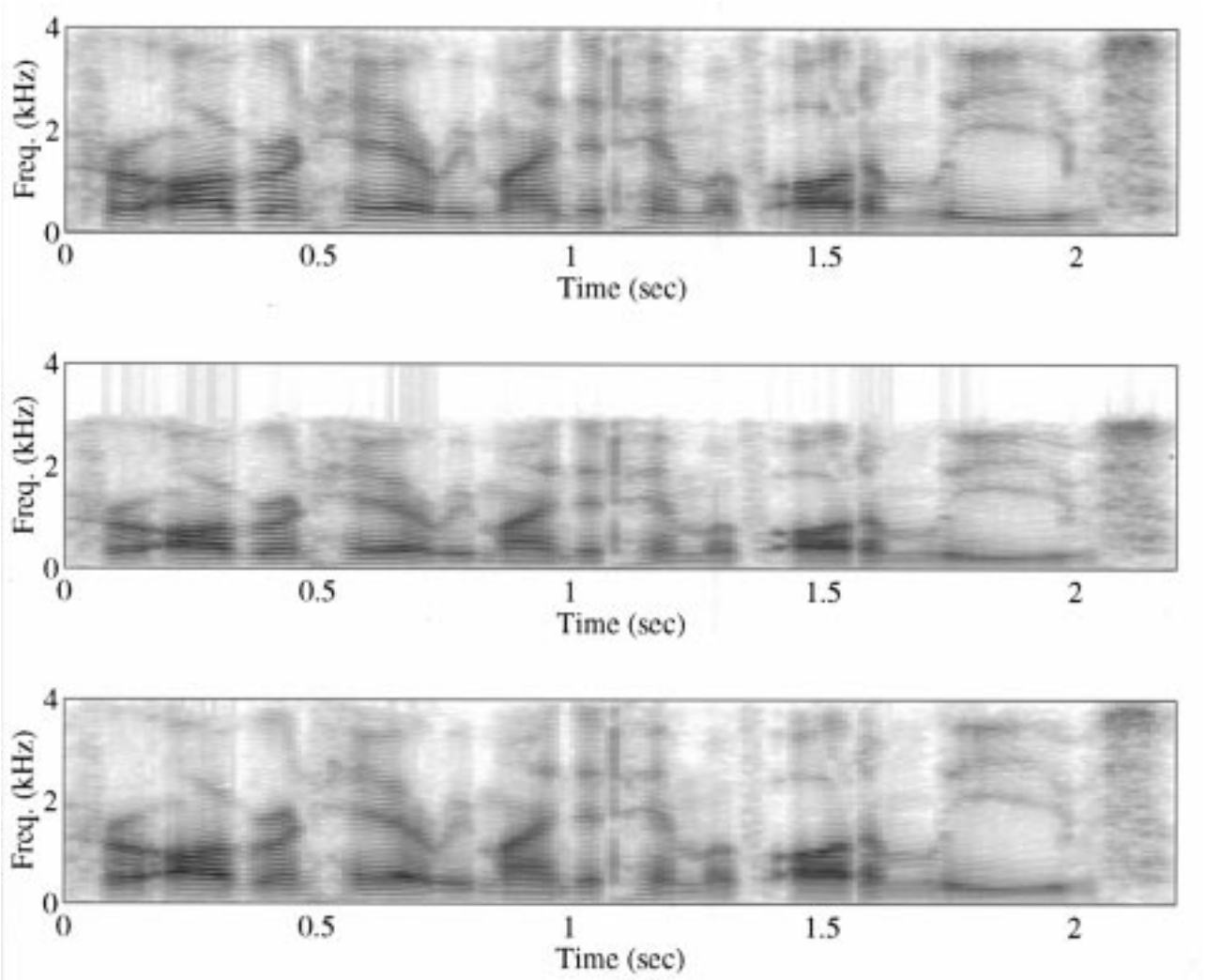

Fig. 7. Spectrographic plots of ABS/OLA frequency- and pitch-scale modification. (a) Original. (b) Frequency-scale modified ($\beta = .75$). (c) Pitch-scale modified by same factor.

human talkers for large pitch changes. The resulting distortion is most notable when reducing the pitch of female talkers; in this case, synthetic speech tends to have an impulsive, "raspy" quality traced to the interpretation of pitch-scale modification as a mixed-phase impulse response convolved with an ideal pulse train [17]. Finally, it should be noted that unlike time-scale modification, pitch-scale modification using the ABS/OLA system is susceptible to pitch doubling and halving errors, due to the lack of self-correction in the phasor interpolation model. Fortunately, pitch detection using sinusoidal model parameters is less prone to gross errors than other methods [28].

IV. Computational Considerations

At first glance, the analysis-by-synthesis algorithm described in Section II-B appears to involve a great deal of computation, since in each analysis frame the five inner product expressions of (9) must be evaluated a total of $M/2+1$ times for each of J sinusoidal components. Furthermore, direct evaluation of the overlap-add synthesis expressions—(2) and (3)—requires the direct computation of J sinusoids at $2N_s+1$ points in each synthesis frame, representing a prohibitive computational load for many real-time applications.

As we will see, however, using the frequency-domain dualities derived in Section II-B1 leads to a formulation of analysis-by-synthesis that operates entirely in terms of discrete Fourier transforms. This frequency-domain formulation of analysis-by-synthesis may then be implemented using the FFT algorithm, significantly improving its computational efficiency. In addition, since constant-amplitude, linear-phase sinusoids are used in overlap-add synthesis, the inverse FFT (IFFT) algorithm may be used there as well. This section discusses techniques for exploiting these observations.

A. Use of the FFT in Analysis-by-Synthesis

The M-point DFT and inverse DFT of an M-point sequence $x[n]$ are defined by

$$X[m] \triangleq \sum_{n=0}^{M-1} x[n]W_M^{mn}, \qquad 0 \le m < M,$$

$$x[n] = \frac{1}{M}\sum_{m=0}^{M-1} X[m]W_M^{-mn}, \qquad 0 \le n < M \quad (24)$$

where $W_M^{mn} = e^{-j(2\pi/M)mn}$. Comparing (24) with (10), if $x[n]$ is nonzero only on the interval $[0, M-1]$, it is easily seen that

$$X[m] \equiv X(e^{j\omega})|_{\omega=(2\pi/M)m}. \tag{25}$$

For the purposes of analysis-by-synthesis, the M-point DFT's of $w_a[n]e_{\ell-1}[n]\sigma[n+kN_s]$ and $w_a[n]\sigma^2[n+kN_s]$ may be expressed as

$$EG_{\ell-1}[m] = \sum_{n=-N_a}^{N_a} w_a[n]e_{\ell-1}[n]\sigma[n+kN_s]W_M^{mn}$$
$$GG[m] = \sum_{n=-N_a}^{N_a} w_a[n]\sigma^2[n+kN_s]W_M^{mn}. \tag{26}$$

Noting that $W_M^{m(n+M)} = W_M^{mn}$, these DFT's may be cast in the form of (24) (provided that $M > 2N_a$) by adding M to the negative summation index values and zero-padding the unused index values.

Recalling the expression for γ_{ij} from (9) for the case when $\omega_\ell = \omega_c[i] = 2i\pi/M$, and making use of the expression for $GG[m]$ given above, it is easily shown that [17]–[19]

$$\gamma_{11} = \tfrac{1}{2}\,\Re e\{GG[0] + GG[2i]\}.$$

Similarly, expressions for γ_{12} and γ_{22} can also be derived as follows:

$$\gamma_{12} = -\tfrac{1}{2}\,\Im m\{GG[2i]\},$$
$$\gamma_{22} = \tfrac{1}{2}\,\Re e\{GG[0] - GG[2i]\}.$$

Examining these relations, it is seen that the first three parameters are determined from the stored values of a single DFT which need only be calculated once per analysis frame. This DFT may be computed via the FFT algorithm using approximately $M\log_2 M$ complex multiplications and additions, yielding dramatic savings in computation over direct evaluation of the inner product terms.

Similar expressions for ψ_1 and ψ_2 may also be derived:

$$\psi_1 = \Re e\{EG_{\ell-1}[i]\}$$
$$\psi_2 = -\Im m\{EG_{\ell-1}[i]\}.$$

These parameters are thus expressed in terms of the stored values of $EG_{\ell-1}[m]$. Of course, since $e_{\ell-1}[n]$ changes for each new component added to the approximation, this DFT must be computed J times per frame. In order to reduce the amount of computation further, the relations derived in Section II-B1 may be used to update $EG_{\ell-1}[m]$.

Combining the results of (13) and (14), the updated "error spectrum" $EG_\ell(e^{j\omega})$ is given by

$$EG_\ell(e^{j\omega}) = EG_{\ell-1}(e^{j\omega}) - \tfrac{1}{2}A_\ell e^{j\phi_\ell}GG(e^{j(\omega-\omega_\ell)}) - \tfrac{1}{2}A_\ell e^{-j\phi_\ell}GG(e^{j(\omega+\omega_\ell)}). \tag{27}$$

Making use of (25), and recalling that $\omega_\ell = 2\pi i_\ell/M$, the updated error DFT $EG_\ell[m]$ is written as

$$EG_\ell[m] = EG_{\ell-1}[m] - \tfrac{1}{2}A_\ell e^{j\phi_\ell}GG[((m-i_\ell))_M] - \tfrac{1}{2}A_\ell e^{-j\phi_\ell}GG[((m+i_\ell))_M] \tag{28}$$

where $((\cdot))_M$ denotes the "modulo M" operator. $EG_\ell[m]$ can therefore be expressed as a simple linear combination of $EG_{\ell-1}[m]$ and circularly shifted versions of $GG[m]$. This method of updating $EG_\ell[m]$ is not only more elegant than that of subtracting successive components from $e_\ell[n]$ and recalculating the DFT, it also represents a considerable improvement in computational efficiency over the direct method. This is because the component $\hat{x}_\ell[n]$ does not have to be evaluated after calculating its parameters, and because the FFT algorithm only has to be used once per analysis frame to calculate $EG_0[m]$; in fact, the only computation required to update $EG_\ell[m]$ is the approximately M additions and multiplications needed to implement (28).

B. Use of the IFFT in Overlap-Add Synthesis

Referring to (2) and using the inverse DFT formula of (24), the expression for $\tilde{s}^k[n]$ may be written as

$$\tilde{s}^k[n] = \sum_{\ell=1}^{J} A_\ell\ \cos(\omega_\ell n + \phi_\ell) = \Re e\left\{\frac{1}{M}\sum_{\ell=1}^{J} MA_\ell e^{j\phi_\ell}W_M^{-i_\ell n}\right\}. \tag{29}$$

From this we see that $\tilde{s}^k[n]$ may be calculated by constructing an M-point sequence in m with values of $MA_\ell e^{j\phi_\ell}$ at $m = i_\ell$ and zero otherwise, then taking the real part of the inverse DFT of this sequence. This establishes a basis for using the IFFT algorithm to perform synthesis.

According to (19), in the presence of time- and frequency-scale modification, a synthetic contribution is given by

$$\tilde{s}^k_{\rho_k,\beta_k}[n] = \sum_{\ell=0}^{J[k]} A_\ell^k \cos(\hat{\omega}_\ell^k n + \zeta_\ell^k) \tag{30}$$

where $\hat{\omega}_\ell^k = \beta_k\ell\omega_o^k + \Delta_\ell^k/\rho_k$ and $\zeta_\ell^k = \phi_\ell^k + \beta_k\ell\omega_o^k\delta^k$. Except for the case when $\beta_k = \rho_k = 1$, the modified frequency terms $\{\hat{\omega}_\ell^k\}$ do not necessarily fall at multiples of $2\pi/M$. However, the IFFT algorithm may still be used to accurately generate modified synthetic contributions by using sinusoids with valid DFT frequencies to approximate components of $\tilde{s}^k_{\rho_k,\beta_k}[n]$.

To accomplish this, it is necessary to specify a variable DFT length $\hat{M}_k$ to use for modification synthesis. Recalling the discussion of approximation accuracy relative to frame length in Section II-B, it is important to adapt $\hat{M}_k$ so that consistent accuracy (as well as consistent computation) is achieved over the varying frame lengths required in fixed and time-varying time-scale modification. To this end, $\hat{M}_k$ may be set to $\hat{M}_k = \nu_s\rho_k N_s$; experiments with audio waveforms sampled at 8–16 kHz indicate that a value of $\nu_s = 5$ is sufficient to guarantee high-quality synthesis over a wide range of modifications.

Given the value of $\hat{M}_k$, each component of $\tilde{s}^k_{\rho_k,\beta_k}[n]$ may be approximated using two sinusoids with valid DFT frequencies adjacent to $\hat{\omega}_\ell^k$. Details of this approximation procedure are presented in [17]–[19]. The parameters of these sinusoids can then be assigned to an $\hat{M}_k$-point sequence $\hat{Z}[m]$ as described previously, and the inverse DFT of $\hat{Z}[m]$ may be calculated by

the IFFT algorithm using on the order of $\hat{M}_k \log_2 \hat{M}_k$ complex multiplications and additions.

V. Discussion

This paper has described the ABS/OLA speech analysis/synthesis system in detail, as well as its application to speech modification. This section presents a comparative discussion of differences between various components of the ABS/OLA system and the STS, summarizes advantages and disadvantages of the ABS/OLA system, and suggests several areas for future research.

Analysis-by-synthesis using an overlap-add model with time-varying gain is an important distinction between the ABS/OLA system and the STS, and has several advantages compared to the peak-picking analysis algorithm used in the STS. Peak-picking assumes spectral magnitude peaks yield optimal sinusoidal component parameters, based on stationarity arguments. By contrast, analysis-by-synthesis attempts to determine the optimal parameters for each component without regard to stationarity. The incorporation of a time-varying gain signal $\sigma[n]$ in the model to account for syllabic signal level variations further reduces the sensitivity of analysis-by-synthesis to stationarity assumptions, particularly when used with the quasiharmonic model formulation. It should be pointed out, however, that for moderate synthesis frame sizes, the time-varying gain signal may be coarsely approximated during synthesis with relatively minor effects on speech quality.

A further distinction of analysis-by-synthesis is its ability to model speech using an indeterminate number of sinusoids. This is in contrast to peak-picking, which limits the number of sinusoids used to the number of identifiable spectral peaks. This ability allows the ABS/OLA system to synthesize very accurate unvoiced speech, without the tonality often associated with peak-picking analysis. Another advantage of analysis-by-synthesis is its ability to deal with sidelobe interference effects. As discussed in Section II-B1, sidelobe interference from component spectra tends to bias sinusoidal model parameters derived from peak-picking. Since analysis-by-synthesis removes each component after determining its parameters, sidelobe effects, which have been observed to produce a slight tonality in synthetic voiced speech using peak-picking analysis, are reduced [17].

To test these observations in a controlled experiment, we used a common overlap-add sinusoidal model with no time-varying gain, and peak-picking and analysis-by-synthesis to determine model parameters. We also formulated hybrid techniques using peak-picking to determine frequency parameters and analysis-by-synthesis to derive amplitude/phase information, and vice-versa. Objective tests using this configuration demonstrate a 5 dB increase in average segmental SNR using analysis-by-synthesis as opposed to peak-picking, primarily due to analysis-by-synthesis generating better estimates of component frequencies [17]. Informal listening tests using the common overlap-add model demonstrate the ability of analysis-by-synthesis to accurately model signals representing perceptually critical transitory events such as plosives, and indicate reductions in the tonality of both voiced and unvoiced synthetic speech.

An important difference between the ABS/OLA system and the STS is that while the STS uses an interpolated parameter model that is not formulated with regard to peak-picking analysis, the synthesis model of the ABS/OLA system maintains the same overlap-add structure used in analysis-by-synthesis. The correlation between analysis and synthesis techniques in the ABS/OLA system leads to consistent and predictable results in speech synthesis and modification.

One of the most significant features of the ABS/OLA system is its ability to perform flexible, high-quality speech modification using an overlap-add sinusoidal model. By manipulating quasiharmonic component frequencies, the refined modification model of the ABS/OLA system manages to preserve waveform shape in the same manner as shape-invariant modification in the STS. Informal listening tests demonstrate that this novel aspect of the ABS/OLA system is critical for performing significant time-scale expansion when compared with a simpler harmonic model overlap-add strategy [17].

A significant computational advantage is gained by using FFT-based overlap-add synthesis directly in the ABS/OLA system. As originally proposed, the STS performs a direct summation of sinusoids generated by oscillators or a lookup table, representing a sizable computational load that varies directly with the number of components in a given signal. By contrast, the overlap-add model used in the ABS/OLA system reduces synthesis computation by using the FFT, and maintains a more consistent level of computation for variable numbers of components. Unfortunately, while ABS/OLA synthesis is computationally efficient, analysis is not. The increased accuracy of analysis-by-synthesis over peak-picking comes at the cost of an order of magnitude increase in computation compared with the simpler peak-picking algorithm, primarily due to the required frequency search [17].

While the ABS/OLA system has proven successful in the applications of speech analysis/synthesis and speech modification, a number of unresolved issues of performance, computation and application remain to be addressed. The computational load of analysis-by-synthesis, while manageable, remains an obstacle for real-time analysis/synthesis. Most of the computational load in analysis-by-synthesis derives from exhaustive frequency searching. While exhaustive searching tends to optimize analysis accuracy, there is sufficient information in speech spectra to allow pruning of the search space with little loss of quality. One strategy that has been suggested is to pick a certain number of peaks from the original spectrum, perform analysis-by-synthesis in narrow bands around those frequencies, then repeat the process after removing the analyzed components.

Stopping conditions in analysis-by-synthesis are defined in terms of how close the approximation error is to zero. While this is useful for analysis of signals with little noise interference, significant additive noise can result in overapproximation and the undesirable tendency to capture noise as well as signal. For speech enhancement applications, it would be helpful to define a different stopping condition based on, for instance, the

"whiteness" of the error for the condition of additive white noise.

While the frequency blanking algorithm for perceptual enhancement of analysis-by-synthesis works well in many cases, it results in poor quality synthesized speech when $w_a[n]$ is too short. Indeed, designing perceptual factors into analysis-by-synthesis is very much an open question. One possible solution is seen by considering Fig. 2. As noted in Section II-C, analysis-by-synthesis using the least-square error norm only accounts for the spectral magnitude at a given candidate frequency; if an error norm were defined which accounts for the match of spectral shape in the mainlobe, then very few spurious components would be chosen, without requiring a hard decision as in frequency blanking.

As noted before, the phasor interpolation algorithm for pitch-scale modification has several advantages; however, speech modified using this approach can sound strange for significant alterations of pitch. To deal with this problem, the pitch modification algorithm requires refinement to account for vocal tract changes that occur in human speech production at different pitch frequencies [31]. A more significant problem is the sensitivity of phasor interpolation to pitch estimation errors and the artifacts that often result. To deal with this problem, the causes of this sensitivity should be isolated and eliminated. An approach currently under investigation is the use of a dynamic-programming based pitch detector to reduce pitch doubling and halving errors that occasionally occur.

Finally, a basic issue in sinusoidal speech modeling is the representation of unvoiced speech or other events that are not naturally modeled using narrowband deterministic processes. While the ABS/OLA system models such events well using only sinusoids, it is not clear that performance in the presence of modification will be uniformly pleasing or natural. Hybrid sinusoidal/stochastic models such as Serra and Smith's *spectral modeling synthesis* system [32] have proven very effective in representing musical sounds; similar techniques applied to speech modeling bear further investigation.

VI. Conclusions

In this paper, a speech analysis/synthesis system was developed based on the combination of an overlap-add sinusoidal model with an analysis-by-synthesis procedure to determine model parameters. The analysis-by-synthesis/overlap-add (ABS/OLA) system is capable of automatically analyzing input speech signals, synthesizing perceptually identical replicas of speech from analyzed parameters, and modifying speech signals to alter their time, frequency, and pitch scale. The ABS/OLA system uses a novel refined quasiharmonic overlap-add model that accounts for waveform evolution in the presence of modifications. This model allows the ABS/OLA system to perform both fixed and time-varying modifications without introducing objectionable artifacts. The phasor interpolation algorithm introduced in the paper represents a novel approach to pitch-scale modification that deals effectively with the problems of bandwidth compression and noise amplification encountered in other pitch-scaling techniques. The computational load required to implement the ABS/OLA system has been significantly reduced by exploiting frequency-domain interpretations of the analysis and synthesis algorithms and their relation to the FFT algorithm. As a result, the ABS/OLA system achieves a combination of modification flexibility, high quality output, and reasonable implementation requirements not found in existing speech analysis/synthesis systems.

References

[1] M. R. Portnoff, "Time-scale modification of speech based on short-time Fourier analysis," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP–30, pp. 374–390, June 1981.

[2] M. R. Schroeder, J. L. Flanagan, and E. A. Lundry, "Bandwidth compression of speech by analytic-signal rooting," *Proc. IEEE*, vol. 55, pp. 396–401, Mar. 1967.

[3] L. D. Braida *et al*., "Matching speech to residual auditory function—a review of past research," ASHA monograph, 1978.

[4] D. H. Klatt, "Review of text-to-speech conversion for English," *J. Acoust. Soc. Amer.*, vol. 82, pp. 737–793, Sept. 1987.

[5] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, pp. 561–580, Apr. 1975.

[6] D. Malah, "Time-domain algorithms for harmonic bandwidth reduction and time scaling of speech signals," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP–27, pp. 121–133, Apr. 1979.

[7] J. L. Flanagan and R. M. Golden, "Phase vocoder," *Bell Syst. Tech. J.*, vol. 45, pp. 1493–1509, 1966.

[8] M. R. Portnoff, "Implementation of the digital phase vocoder using the fast Fourier transform," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP–24, pp. 243–248, June 1976.

[9] 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. 236–242, 1984.

[10] L. B. Almeida and J. M. Tribolet, "Nonstationary spectral modeling of voiced speech," *IEEE Trans. Acoust., Speech and Signal Processing*, vol. ASSP–31, pp. 374–390, June 1983.

[11] P. Hedelin, "A tone-oriented voice-excited vocoder," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Mar. 1981, pp. 205–208.

[12] 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.

[13] T. F. Quatieri and R. J. McAulay, "Speech transformations based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP–34, pp. 1449–1464, Dec. 1986.

[14] ———, "Phase coherence in speech reconstruction for enhancement and coding applications," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, May 1989, pp. 207–210.

[15] ———, "Shape invariant time-scale and pitch modification of speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP–40, pp. 497–510, Mar. 1992.

[16] R. J. McAulay and T. F. Quatieri, "Computationally efficient sine-wave synthesis and its application to sinusoidal transform coding," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Apr. 1988, pp. 370–373.

[17] E. B. George, "An analysis-by-synthesis approach to sinusoidal modeling applied to speech and music signal processing," Ph.D. dissertation, Georgia Inst. Technol., Atlanta, GA, 1991.

[18] 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, June 1992.

[19] ———, "Audio analysis/synthesis system," U.S. Patent 5 327 518, July 1994.

[20] J. S. Marques *et al.*, "Harmonic coding of speech at 4.8 kb/s," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Apr. 1990, pp. 17–20.

[21] D. W. Griffin and J. S. Lim, "Multiband excitation vocoder," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP–36, pp. 1223–1235, Aug. 1988.

[22] A. Spanias, "Speech coding: A tutorial review," *Proc. IEEE*, vol. 82, pp. 1541–1582, Oct. 1994.

[23] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Englewood Cliffs, NJ: Prentice-Hall, 1978.

[24] E. B. George and M. J. T. Smith, "A new speech coding model based on a least-squares sinusoidal representation," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Apr. 1987, pp. 1641–1644.

[25] J. S. Marques and L. B. Almeida, "A background for sinusoid based representation of voiced speech," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, Apr. 1986, pp. 1233–1236.
[26] F. J. Harris, "On the use of windows for harmonic analysis with the discrete fourier transform," *Proc. IEEE*, vol. 66, pp. 51–83, Jan. 1978.
[27] S. R. Quackenbush, T. P. Barnwell, III, and M. A. Clements, *Objective Measures of Speech Quality*. Englewood Cliffs, NJ: Prentice-Hall, 1988.
[28] R. J. McAulay and T. F. Quatieri, "Pitch estimation and voicing detection based on a sinusoidal speech model," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Apr. 1990, pp. 249–252.
[29] T. F. Quatieri and R. J. McAulay, "Mixed-phase deconvolution of speech based on a sine-wave model," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Apr. 1987, pp. 649–652.
[30] R. J. McAulay and T. F. Quatieri, "Phase modeling and its application to sinusoidal transform coding," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Apr. 1986, pp. 1713–1715.
[31] S. Seneff, "System to independently modify excitation and/or spectrum of speech waveform without explicit pitch extraction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP–30, pp. 566–578, Aug. 1982.
[32] X. Serra and J. O. Smith III, "Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition," *Comput. Music J.*, vol. 14, pp. 12–24, Winter 1990.

E. Bryan George (S'86–M'90) received the B.E.E. degree in 1985 and the Ph.D. degree in 1991, both from the Georgia Institute of Technology.

After completing doctoral work, he joined Sanders, a Lockheed-Martin Company, Nashua, NH, where he performed research in speech enhancement and corpus development for defense applications. In 1994, he became a Member of Technical Staff, Corporate Research and Development, Texas Instruments, Dallas, TX, where he has performed research in low bit rate speech coding and analysis-based music and singing voice synthesis.

Dr. George has been active as a reviewer for the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING, a member of the Signal Processing Society Audio Technical Committee, Publicity Chair for the Workshop on Applications of Signal Processing to Audio and Acoustics, and as a Technical Program Committee member for the Workshop on Multimedia Signal Processing.

Mark J. T. Smith (S'82–M'84–SM'90–F'95) received the S.B. degree from the Massachusetts Institute of Technology, Cambridge, in 1978, and the M.S. and Ph.D. degrees from the Georgia Institute of Technology, Atlanta, in 1979 and 1984, respectively, all in electrical engineering.

He is presently a Professor in the School of Electrical and Computer Engineering at Georgia Tech, where he is involved in research in the areas of speech and image processing, filterbanks and wavelets, and object detection and recognition.

Dr. Smith has authored many papers in the area of signal processing, four of which have received IEEE awards. He is the co-author (with R. Mersereau) of two introductory books entitled *Introduction to Digital Signal Processing* (Wiley, 1992) and *Digital Filtering* (Wiley, 1994). He is also co-editor (with A. Akansu) of *Wavelets and Subband Transforms: Design and Applications* (Kluwer) and the co-author of *A Study Guide to Image Processing* (Scientific, 1997). He is a past Chairman of the IEEE SP Digital Signal Processing Technical Committee, a member of the IEEE SP Multimedia Technical Committee, and serves on the Board of Governors of the IEEE Signal Processing Society. He has served as an Associate Editor for the IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING, and as a member of the MIP's Advisory Board of the National Science Foundation. He has been active as a member of the Organizing Committees for the IEEE DSP Workshops since 1988, serving as General Chairman in 1992 and 1994. He has also been active on the organizing committees for the SPIE Visual Communications and Image Processing Conferences (VCIP) since 1990, most recently serving as General Co-Chairman in 1996.