PRACTICAL HIGH-QUALITY SPEECH AND VOICE SYNTHESIS USING FIXED FRAME RATE ABS/OLA SINUSOIDAL MODELING

E. Bryan George

DSP Solutions Research and Development Center
Texas Instruments Incorporated
P. O. Box 655303, M/S 8374
Dallas, Texas 75265 USA
ebg@ti.com

ABSTRACT

This paper describes algorithms developed to apply the *Analysis-by-Synthesis/Overlap-Add* (ABS/OLA) sinusoidal modeling system to real-time speech and singing voice synthesis. As originally proposed, the ABS/OLA system is limited to unidirectional time-scaling, and relies on variable frame length to accomplish time-scale modification. For speech and voice synthesis applications, unidirectional time scaling makes effective looping to produce sustained vocal sounds difficult, and variable frame length makes real-time polyphonic synthesis problematic. This paper presents a reformulation of the basic ABS/OLA system to deal with these issues, which is termed *Fixed-Rate ABS/OLA* (ABS/OLA-FR).

1. INTRODUCTION

Music and speech synthesis technologies are finding increasing use in applications requiring audio and voice interfaces. As the need for these synthesis technologies has grown, their limitations have become apparent. Music synthesis, while capable of producing high-quality, expressive musical instrument sounds, is relatively incapable of replicating the most expressive instrument known, the singing voice.

Conversely, speech synthesis, while impressive from the standpoint of computational linguistics, has historically been limited in terms of subjective signal quality, expressivity, musicality, and similar "production values." These limitations remain a significant barrier to the application of speech synthesis in multimedia and entertainment applications.

Recently, a set of speech synthesis systems based on the concatenation of appropriately modified speech waveforms has emerged [1]-[4]. Such "data-driven" systems have shown considerable potential for high-quality speech synthesis, since they begin with unit inventories of natural speech, rather than highly abstracted speech production models.

The potential of concatenative speech synthesis is dramatically demonstrated by the LYRICOS singing voice synthesizer [5]. This MIDI file-driven synthesizer combines the positive aspects of music and speech synthesizers, producing synthetic singing voice with a very natural and expressive character.

2. THE ABS/OLA SINUSOIDAL MODEL

The quality of a concatenative speech synthesizer is limited by the ability of its underlying waveform synthesis engine to flexibly modify the pitch, duration, and other perceptual characteristics of speech units without producing noticeable artifacts. The synthesis model used in the LYRICOS system is the *Analysis-by-Synthesis/Overlap-Add* (ABS/OLA) system [6]-[8], which has been demonstrated to produce very high quality speech modifications at a moderate computational cost.

The ABS/OLA system uses an overlap-add sinusoidal model of the form

$$s[n] = \sigma[n] \sum_k w_s[n - kN_s] s^k[n - kN_s] \qquad (1)$$

where $\sigma[n]$ represents the slowly varying *envelope* of the speech unit. The overlapping signals $s^k[n]$ con-

tributing to *synthesis frame* k are sums of constant-amplitude, constant-frequency sinusoids generated via the inverse FFT (IFFT) algorithm.

Each contribution $s^k[n]$ has the general form

$$s^k[n] = \sum_i A_i^k \cos(\omega_i^k n + \phi_i^k) \qquad (2)$$

where the frequencies ω_i^k are quasi-harmonically related, and where the complementary *synthesis window* $w_s[n]$ controls fading in the overlap-add model between contributions from frame to frame. The overlap-add process is illustrated in Figure 1. Note that analysis parameter values are synchronized with synthesis frame boundaries, *i.e.* parameters $\{A_i^k, \omega_i^k, \phi_i^k\}$ contribute to frame $k-1$ as well as frame k.

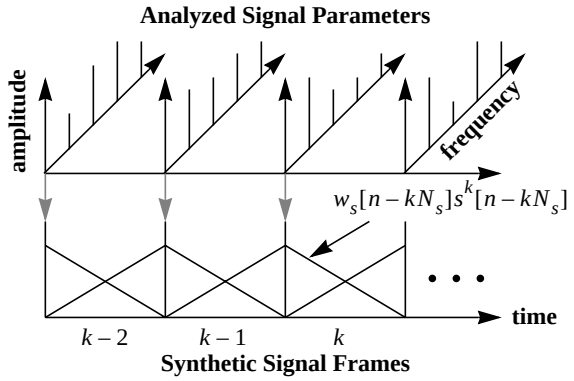

Figure 1. Illustration of overlap-add synthesis in the ABS/OLA sinusoidal model.

As originally proposed, the ABS/OLA system performs time-scale modification by accessing analysis parameters at a fixed frame rate and synthesizing with a variable frame rate. This synthesis structure is illustrated in Figure 2. In this figure, scale factors ρ_k control the length of successive synthesis frames by multiplication, *i.e.* a factor of 2 produces a synthesis frame with twice the length of the original, corresponding to a "slow down" by the same factor.

While variable frame rate synthesis effectively produces time-scale modification, it introduces two serious difficulties for practical implementation in a real-time polyphonic voice synthesizer. First, since the synthesis frame length varies depending on the time-scale factor, and since time-scale factors may differ independently among several voices, polyphonic synthesis requires an independent IFFT for each voice. Furthermore, to insure accuracy each IFFT must vary in length proportionally with synthesis frame length, the synthesis window length must vary dynamically as well, and variable frame lengths require a complicated buffer management strategy to properly mix voices together before output.

From an implementation standpoint, variable frame rate synthesis is undesirable, since the duplicate IFFT and windowing operations for each new voice, together with the difficulties of buffer management, make the synthesizer unnecessarily complicated. Also note that ρ_k is required to be positive, since synthesizing a negative-length frame is impossible. Time-scale modification in this model must therefore be *unidirectional*, seriously limiting the looping strategies available to the ABS/OLA system.

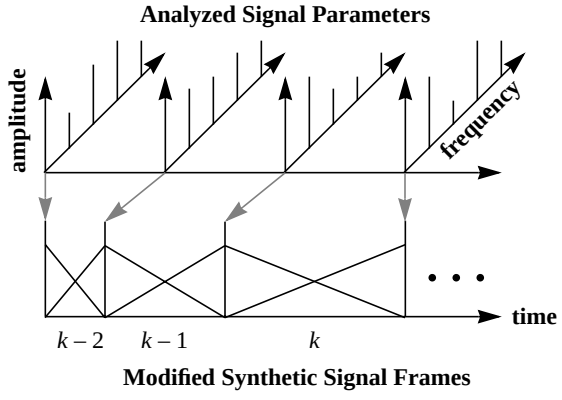

Figure 2. Variable frame rate ABS/OLA synthesis with time-scale factors $\{\rho_{k-2}, \rho_{k-1}, \rho_k\} = \{.5, 1, 1.5\}$.

3. FIXED FRAME-RATE ABS/OLA SYNTHESIS

Since the described implementation problems result entirely from variable synthesis frame length, the question arises whether it is possible to formulate a synthesis strategy with fixed frame lengths, using a different mechanism to implement time-scale modification.

The goal of time-scale modification is to produce a change in the time evolution of the synthesized signal. If the output frame rate is fixed, producing time-scale modification requires that we change the access

rate of analysis parameters used to synthesize each frame.

Such a process is seen in Figure 3, where time-scale factors $\hat{\rho}_k$ now control the *analysis data access point*, rather than synthesis frame length, in the following sense: If analysis parameters used to synthesize frame k are extracted from the access point n_k (in the original time scale), then the analysis parameters used to synthesize frame $k+1$ are extracted from the access point given by

$$n_{k+1} = n_k + \hat{\rho}_k N_s$$

In this context, $\hat{\rho}_k$ has an opposite effect from ρ_k, *i.e.* $\hat{\rho}_k = 2$ implies that analysis parameters are being accessed twice as fast as normal, implying a 2x "speed up."

Algorithmically, synthesis using this modification strategy is practically identical to traditional ABS/OLA synthesis once suitable analysis parameters have been extracted. The lone exception is computation of the frame-dependent time shift δ_k used to guarantee proper inter-frame coherence when modified contributions are summed together [7],[8]. In ABS/OLA-FR, the recursion for δ_k becomes

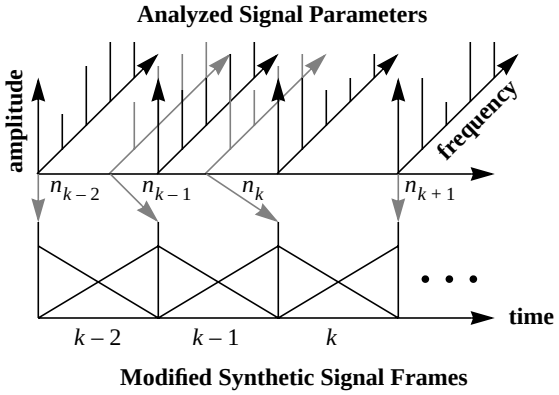

Figure 3. Fixed frame rate ABS/OLA synthesis with time-scale factors $\{\hat{\rho}_{k-2}, \hat{\rho}_{k-1}, \hat{\rho}_k\} = \{.6, .8, 1.6\}$.

$$\delta_{k+1} = \frac{\omega_o^k\left(\beta_k\delta_k + (\beta_k - \hat{\rho}_k)\frac{N_s}{2}\right) + \omega_o^{k+1}\left((\beta_{k+1} - \hat{\rho}_k)\frac{N_s}{2}\right)}{\beta_{k+1}\omega_o^{k+1}}$$

where ω_o^k is the radian fundamental frequency of each contribution, and where β_k is the multiplicative frequency- (or pitch-) scale modification factor associated with the k-th synthetic contribution.

Figure 4 shows a block diagram of polyphonic synthesis using a fixed frame-rate approach to ABS/OLA synthesis, which we term *fixed-rate ABS/OLA* (ABS/OLA-FR). With a common frame rate now used for multiple voices, we may apply different time- and frequency-scale factors to each voice, and linearity of the IFFT implies that parameters for each voice may be mixed in the frequency domain, then input to a single IFFT block. This implies less computation than variable frame rate synthesis, as well as a more consistent level of computation as a function of polyphony. Furthermore, since time-scale modification is now based on relative analysis data access points rather than synthesis frame length, time-scale modification in ABS/OLA-FR may be *bidirectional*, allowing for more flexible looping strategies.

3.1 Parameter Interpolation

Although the implementation advantages of fixed frame-rate synthesis are clear, fixed-rate synthesis introduces a difficult problem: Analysis data do not necessarily exist at the random access points specified by time-scale modification in ABS/OLA-FR. The simplest method for dealing with this problem is to perform analysis at a higher frame rate, then constrain time-scale modification such that access is limited to times where analysis data exist. This, of course, is unsatisfactory, since it increases the amount of data storage required for analysis parameters and/or severely limits the flexibility of time-scale modification.

Fortunately, the functional, quasi-harmonic form of ABS/OLA synthesis seen in Equations 1 and 2 provides the means to extract analysis parameters at any time index. To see this, consider Equation 1, where the synthesis window $w_s[n]$ has support limited to two consequtive synthesis frames, and where only the i-th component of Equation 2 is considered:

$$\begin{aligned} s[n + kN_s] = {} & a_1^k[n]\cos(\omega_j^k n + \phi_j^k) \\ & + a_2^k[n]\cos(\omega_j^{k+1} n + \phi_j^{k+1}) \end{aligned} \quad (3)$$

for $0 \le n < N_s$, where

$$a_1^k[n] = A_j^k \sigma[n + kN_s] w_s[n]$$
$$a_2^k[n] = A_j^{k+1} \sigma[n + kN_s] w_s[n - N_s]. \quad (4)$$

By considering the time-varying gain to be piecewise linear between sample points, and by using a synthesis window with a differentiable functional form, it is possible to substitute the continuous variable t for the discrete variable n in the above equations. The trigo-

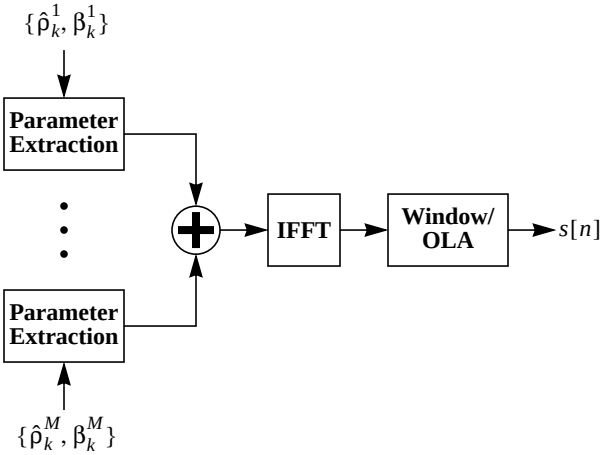

Figure 4. Block diagram of M-voice polyphonic synthesis in the ABS/OLA-FR system.

nometric form of Equation 3 then allows us to express $s[n + kN_s]$ as a complex *analytic signal* of the form

$$\hat{x}(t) = A(t)e^{j\Phi(t)} = \alpha(t) + j\beta(t)$$

for $0 \le t < N_s$. The amplitude, frequency and phase parameters for *any* value of t (including non-integer values) may be derived by solving for the instantaneous amplitude, frequency, and phase of $\hat{x}(t)$ using the following relations:

$$A(t) = \sqrt{\alpha^2(t) + \beta^2(t)}$$
$$\omega(t) = \frac{\alpha(t)\beta'(t) - \beta(t)\alpha'(t)}{A^2(t)}$$
$$\phi(t) = \text{atan}(\beta(t)/\alpha(t)).$$

4. CONCLUSION

This paper has described a number of algorithm developments required to implemented the Analysis-by-Synthesis/Overlap-Add (ABS/OLA) sinusoidal modeling system in a practical, MIDI-driven real-time polyphonic music synthesizer. Our new formulation, termed fixed-rate ABS/OLA (ABS/OLA-FR), uses a fixed frame-rate and fixed FFT length to enable a single, efficient synthesis engine to produce multiple voices.

5. REFERENCES

[1] E. Moulines and F. Charpentier, "Pitch synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Comm.*, 9:453-467, Dec. 1990.

[2] T. Dutoit, "High quality text-to-speech synthesis: A comparison of four candidate algorithms," in *Proc. ICASSP 94*, pp. I-565 - I-568, Apr. 1994.

[3] M. W. Macon and M. A. Clements, "Speech concatenation and synthesis using an overlap-add sinusoidal model," in *Proc. ICASSP 96*, pp. 361-364.

[4] X. Huang et al, "Recent improvements on microsoft's trainable text-to-speech system - WHISTLER," in *Proc. ICASSP 97*, pp. 959-962.

[5] M. W. Macon et al, "A singing voice synthesis system based on sinusoidal modeling," in *Proc. ICASSP 97*, pp. 435-438.

[6] E. B. George and M. J. T. Smith, "Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model," *IEEE Trans. on Speech and Audio*, 5(5):389-406.

[7] 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.*, 40(6):497-516, June 1992.

[8] E. B. George and M. J. T. Smith, "Audio Analysis/Synthesis System," United States Patent #5,327,518, July 5, 1994.