AN ENHANCED ABS/OLA SINUSOIDAL MODEL FOR WAVEFORM SYNTHESIS IN TTS

Michael W. Macon [1*] *Mark A. Clements* [2]

[1]Dept. of ECE, Oregon Graduate Institute, Portland, OR 97291-1000
[2]Dept. of ECE, Georgia Institute of Technology, Atlanta, GA 30332

ABSTRACT

This paper describes a method for text-to-speech waveform synthesis based on the Analysis-by-Synthesis/Overlap-Add (ABS/OLA) sinusoidal model. This model has been shown in previous work to be a useful framework for pitch and time-scale modification of both speech and music signals. This paper explores some extensions of the original ABS/OLA formulation that attempt to overcome specific artifacts, including a phase dithering approach for unvoiced speech synthesis and an improved pitch modification method that compensates for undesirable energy modulation effects. The implementation of the model within a text-to-speech synthesis (TTS) system is described, and the results of a listener evaluation of the method are discussed.

1. THE ABS/OLA SINUSOIDAL MODEL

The Analysis-by-Synthesis Overlap-Add (ABS/OLA) sinusoidal model [1, 2], represents an input signal $s[n]$ by a sum of equal-length, overlapping short-time signal frames $s_k[n]$.

$$s[n] \approx g[n] \sum_k w[n - kN_s] s_k[n] \qquad (1)$$

where N_s is the frame length, $w[n]$ is a tapered window function, $g[n]$ is a slowly time-varying gain envelope, and $s_k[n]$ represents the kth frame of the synthesized signal. Each term $s_k[n]$ is represented as the sum of a small number of constant-frequency, constant-amplitude sinusoidal components, given by

$$s_k[n] = \sum_{l=0}^{L-1} A_l^k \cos(\omega_l^k n + \phi_l^k) \qquad (2)$$

where L is the number of sinusoidal components in the frame, and A_l^k, ω_l^k, and ϕ_l^k are the sinusoidal amplitudes, frequencies, and phases, respectively.

An iterative analysis-by-synthesis procedure is performed to find the 'optimal' parameters for each signal frame, based on a mean-squared error criterion. This iterative search can be viewed as an example of a *matching pursuit* algorithm [3] for expanding a signal over an overcomplete basis set. The basis vectors in the set are cosines at frequencies in each bin of a DFT of the analysis frame, weighted adaptively by the gain envelope $g[n]$. The final frequencies $\{\omega_l\}$ chosen for a particular frame are not constrained to be exact multiples of a fundamental. However, only a single component with the largest amplitude near each harmonic of F_0 is kept in the representation—this is referred to as a 'quasiharmonic' representation.

Synthesis is performed by a constant frame-rate (not pitch synchronous) overlap-add procedure that uses the inverse fast Fourier transform to compute each term $s_k[n]$, rather than sets of oscillator functions, as in other sinusoidal models [4, 5]. Time-scale modification is achieved by changing the time evolution rate of the model parameters for each frame and changing the frame duration, while imposing phase constraints on the sinusoidal components to maintain general waveform shape characteristics over the frame. Pitch modification is performed within this same context by altering the component frequencies, phases, and amplitudes in such a way that the fundamental frequency is modified while the general spectral shape is maintained [2].

1.1. Application to TTS

In previous work, we described the incorporation of ABS/OLA into systems for concatenative speech synthesis [6, 7] and singing voice synthesis [8]. In this approach, the unit inventory of the synthesizer is stored as sets of ABS/OLA model parameters instead of as waveforms. To synthesize a new utterance, the model parameters for the sequence of concatenated units are extracted from the inventory and used to generate the pitch- and duration-scaled speech. Modification of a single, continuous utterance is somewhat easier than concatenating dissimilar segments. Several issues become more critical in the TTS application, including the following:

Pitch epoch estimation To perform F_0 and duration modification with ABS/OLA, it is necessary to find an "anchor point" in each frame so that it can be properly aligned to the previous frame after modification. This is accomplished with the 'pitch pulse onset time' estimator described in [4], which uses a correlation-like measure to find the glottal closure instant. It has been our experience that this algorithm often makes errors. One nice feature of ABS/OLA is that the pitch pulse onset time estimates become irrelevant for resynthesis of continuous speech when little or no modification of the pitch or duration is desired. In this default case, the original waveform is reconstructed perfectly because the interrelationship of adjacent frames is taken into account in the alignment. However, at concatenated unit boundaries, this onset time estimate is much more critical, since there is no time relationship of frames across the join point. Poor alignment can cause a garbled speech quality. To combat this problem, we have found it advantageous to utilize a set of pitchmarks derived from an electroglottograph recording to 'seed' the onset determination algorithm. We have also employed

*This work was supported in part by grants from Lockheed-Martin, Texas Instruments, Intel Corporation, and by National Science Foundation grant IIS-9875950.

a smoothing algorithm that identifies and corrects gross errors in the onset times [7].

Duration modification The current implementation of the synthesizer sits within the Festival TTS system [9]. One advantage of the Festival architecture is that time-stamped information about all levels of the linguistic description of the utterance can be retrieved during waveform synthesis (e.g., to what phoneme/syllable/word does frame n belong?). This information can be used, for example, to smooth concatenation discontinuities in vowels but avoid smoothing in plosives. Another key use of the linguistic information from the TTS system is in duration modification of stops and segments bordering silence. In these cases, the duration can be lengthened by inserting silence in the stop closure, instead of uniformly stretching the burst and closure.

The duration of a signal is expanded in ABS/OLA synthesis by lengthening each synthesis frame by the desired time stretch factor. For moderate duration stretch factors, this approach is adequate. However in some applications (e.g., singing synthesis [8]), a single segment may need to be lengthened by a much greater factor, and it becomes undesirable to compute very long synthesis frames. In these cases, we have employed a looping strategy that repeats a few frames from the center of the segment (usually a vowel). A better approach, reported in [10], might be to use a reformulated OLA synthesis, where a frequency-domain interpolation of the sinusoidal parameters is used. We have not yet explored this approach for TTS.

The results of a listener comparison of ABS/OLA to a time-domain, pitch-synchronous method are described in Section 3..

2. EXTENSIONS TO ABS/OLA

This section describes some extensions of the ABS/OLA model, designed with the goal of overcoming artifacts that commonly arise in pitch- and time-scaling algorithms.

2.1. Phase randomization

A commonly-cited problem in both sinusoidal and time-domain speech modification algorithms is the existence of so-called 'tonal noise' artifacts in unvoiced or partially-voiced speech after time-scale expansion. It is likely that this artifact is perceived because long-term correlations are introduced into previously 'random' segments, and the human auditory pitch detection mechanism begins to recognize a periodicity. This suggests that the perception of randomness can be maintained by disrupting these long-term periodicities.

Several researchers have proposed harmonic/stochastic decompositions of the signal for coding and modification (e.g., [5, 11]). Most of these are based on representing the periodic portion of the signal by a sinusoidal model and then modeling the residual as the output of a time-varying filter excited by white noise. This is often effective, but can sometimes result in the harmonic and noise parts being perceived as two distinct sources by the listener, rather than as a single, 'fused' source.

In [12], we proposed an extension to ABS/OLA that attempts to disrupt undesired periodicities by performing a subdivision of the synthesis frames and introducing random phase shifts to some of the sinusoids. It can be shown [12] that gradually increasing the magnitude of this random shift causes each component to be transformed gradually into a narrowband noise with the same power as the original sine wave, while at the same time keeping the original overlap-add computational framework. In earlier work by McAulay and Quatieri [13], a related technique applied to an oscillator-type sinusoidal model was referred to as 'phase dithering.'

The first of two audio examples included on the conference CD-ROM demonstrates the use of this phase randomization technique on a female utterance of the words "cyclical programs," with an exaggerated time stretch. The three utterances in the file are (1) *original*, (2) *time-stretched*, and (3) *time-stretched with phase randomization*. Results of a listener subjective evaluation are described in Section 3..

2.2. Pitch modulation compensation

When pitch modification is used to lower F_0, it is common for a kind of 'choppy' or 'pulsy' modulation artifact to arise when using either time-domain or frequency-domain approaches. This artifact is most pronounced in sounds like voiced or unvoiced fricatives. In TD-PSOLA [14], it is straightforward to understand why this modulation occurs—when pitch pulses in a sequence are windowed, the sum of the analysis windows is no longer 1.0 when the windows are moved further apart and re-summed.

It is at first unclear why ABS/OLA and other sinusoidal models should also exhibit this behavior, since their mechanism for pitch modification operates in the frequency domain. As shown in the left panels of Figure 1, a smooth amplitude envelope representing the vocal tract resonances is found by a convolution with a function $W(\omega)$. This interpolation function typically has the property that $W(0) = 1.0$ and $W(\omega) = 0$ for all $|\omega| < \omega_0$. The smooth envelope is then resampled at a new fundamental frequency $\hat{F}_0$ to maintain the original formant pattern but change the pitch.

However, if the time-domain equivalent of this frequency-domain operation is considered, it can be seen that a time-modulation artifact still arises. This is depicted on the right side of Figure 1. The time domain equivalent to $W(\omega)$ has zero crossings at multiples of T_0, and the spectral resampling operation corresponds to a periodic replication at multiples of the new period $\hat{T}_0 = 1/\hat{F}_0$. When the fundamental frequency is lowered, these periodic replications will not sum to a constant, but rather will exhibit a time oscillation with a frequency $\hat{F}_0$.

The resulting modulation term can be shown to be

$$c[n] = \beta + 2\beta \sum_{k=1}^{K} W\left(k\beta\omega_0\right) \cos\left(k\beta\omega_0 n\right), \qquad (3)$$

where β is the pitch modification factor ($\beta < 1.0$ lowers F_0) and $K = \lfloor \frac{1}{\beta} \rfloor$. For $\beta > 1.0$, $c[n]$ causes a simple gain adjustment; for $0.5 < \beta < 1.0$, $c[n]$ contains a cosine term that creates the amplitude modulation

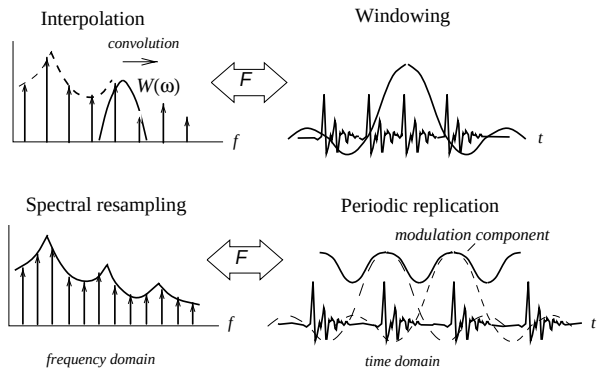

Figure 1. Pitch modification (lowering) involves (upper left) interpolation in the frequency domain using a window function with zero crossings at multiples of F_0 and (lower left) resampling at multiples of a new fundamental $\hat{F}_0$. In the time-domain, this corresponds to (upper right) windowing with a time-domain window with zero crossings at multiples of T_0 and (lower right) periodic replication at the new period $\hat{T}_0$. The sum of these shifted time-domain equivalents to the interpolation function $W(\omega)$ creates an amplitude modulation.

described above. Reference [7] contains additional details of the derivation.

Given this analytic description of the expected modulation term, a simple approach to alleviate it is to simply divide by $c[n]$ to produce a new output signal $s'[n] = s[n]/c[n]$. (In practice, limiting $c[n]$ to a minimum value is necessary when $\beta \leq 0.5$, to avoid zeros in $c[n]$.) An illustration of the effect of this compensation is shown in Figure 2.

The second of two audio examples on the conference CD-ROM demonstrates the use of this compensation technique on a male utterance of the word "splurged" after pitch modification. The three utterances in the file are (1) *original*, (2) *pitch-lowered*, and (3) *pitch-lowered with modulation compensation*.

3. EVALUATION

ABS/OLA TTS In order to evaluate the quality of speech produced by the ABS/OLA-based text-to-speech system, a listener evaluation of the algorithm in comparison to an implementation of a pitch-synchronous time-domain method (similar to [14]) was performed. The ABS/OLA extensions described in Section 2 were not utilized in this test.

Twenty-five subjects were asked to compare 30 pairs of sentences in a randomized A/B comparison. Synthesis units to be concatenated were selected from an inventory of continuous speech based on the similarity of their linguistic context to the target. All text/linguistic analysis modules were the same for each case. The text items used as input to the synthesizers were a set of short declarative sentences (from the "Harvard" sentences). For each trial, the text representation of the sentence was displayed for the subject, and the two synthesized sentences were presented via headphones. Each subject was asked to select a preference "in terms of overall sound quality." Across all subjects and test cases, the results were as follows:

Prefer sinusoidal model	52 %
Prefer time-domain method	48 %

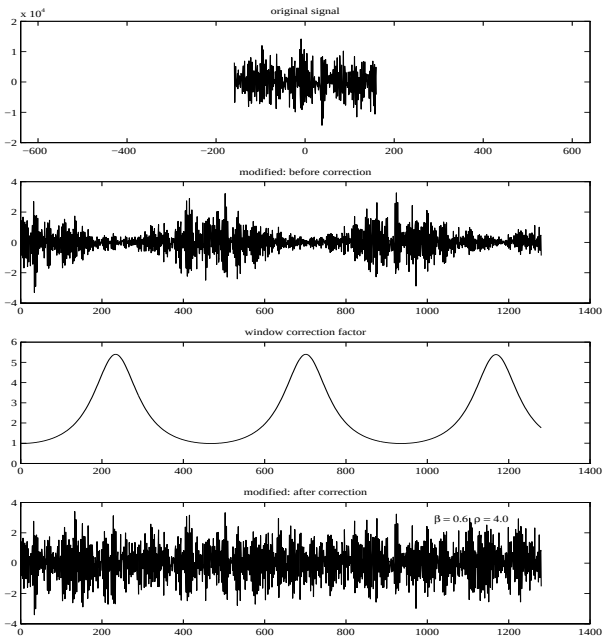

Figure 2. Compensation technique applied to an unvoiced speech segment. From top to bottom (*i*) original speech, (*ii*) pitch modified and time expanded signal exhibiting modulations, (*iii*) correction factor $1/c[n]$, and (*iv*) enhanced speech.

Although a slight preference for the sinusoidal model was shown, this result is not statistically significant. (Based on 750 trials, the probability that this result is due to chance is 0.2.) Taken as a whole, this pool of listeners did not prefer one algorithm over the other. It should be emphasized that the algorithms were tested as part of a full TTS system with several interdependent modules. Upon review of the synthesized audio files, it was clear that the synthesis results for each sentence were either both very good or both very poor in terms of naturalness, depending to a large extent on the set of concatenated units selected from the inventory during unit selection. Further details concerning the test can be found in [7].

Phase randomization To confirm the appropriateness of the phase randomization approach, another subjective comparison test was conducted using the same group of subjects. The subjects were asked to compare 32 pairs of utterances in a blind A/B test, where each pair consisted of one utterance synthesized with the phase randomization algorithm applied and one synthesized using ABS/OLA without this extension. The speech material used as input to the algorithm consisted of eight short phrases containing several unvoiced phonemes, spoken by male and female voices. The sinusoidal model analysis procedure was run on each of the sentences, and a voicing decision was made in each frame. Additional constraints were applied to prevent incorrect voicing decisions in glottal onsets and other voiced transient signal segments.

Four test conditions were applied to each of the eight sentences. Time-scale modifications by factors of $\rho = 2.0$, 3.0, and 4.0 (slower speech) were applied with no pitch modification, and time-scale modifica-

tion by a factor of 3.0 was also applied in combination with a pitch modification by a factor of $\beta = 1.5$ (higher pitch).

The results of the four test conditions were as follows

test	modification factors	% prefer phase rand
A	$\beta = 1.0, \rho = 2.0$	81.0
B	$\beta = 1.0, \rho = 3.0$	79.0
C	$\beta = 1.0, \rho = 4.0$	73.5
D	$\beta = 1.5, \rho = 3.0$	72.5

Each value given represents a percentage of responses preferring the phase randomization method over the standard modification method, averaged over the eight utterances and 25 test subjects. Based on this number of trials, the test results show a preference for the phase randomization method that is statistically significant ($p < 0.001$) in all cases.

4. DISCUSSION

As shown, the proposed extensions to ABS/OLA were shown to improve quality in isolated tests, especially for the handling of unvoiced speech. However, for general use in the TTS system, the benefit of these extensions is counterbalanced with the problems of mis-estimation of other parameters. For example, errors in F_0 estimation can lead to incorrect judgements of the degree of voicing, causing the phase randomization approach to be used in inappropriate locations.

All speech modification methods depend critically on a few (so-called 'solved') problems: (*i*) accurate F_0 tracking, (*ii*) glottal epoch detection (or equivalently, phase unwrapping [5]), and (*iii*) accurate voicing decisions. For high-quality TTS applications, a single error in these parameters can cause an audible and objectionable distortion in the output speech. Our future efforts will focus on improving these 'building blocks,' as well as the ability of the algorithm to robustly handle errors in these parameters.

Based on the results of our first test, is not yet clear whether the ABS/OLA model carries significant advantages over simpler methods like TD-PSOLA, especially given its added computational expense. In We believe that these initial results should be considered with caution, however, since others (e.g., [15]) have shown a preference for sinusoidal models over TD-PSOLA in listener evaluations.

It also may be true that the test we conducted was too limited able to demonstrate the advantages of the ABS/OLA technique. The TTS system used speech units selected to match the desired F_0 and duration. In cases where the synthesis results were of very high quality, the units selected were matched to the desired context very well. Thus the responsibilities placed on the prosody modification algorithm in these cases were very slight (pitch modification factors close to 1.0.). As would be expected, the superiority of one algorithm over the other was thus less apparent, since both were essentially resynthesizing the unmodified speech. We plan to systematically compare these algorithms under more severe prosodic modifications in the future, and also repeat the tests for specific voices (like breathy females) for which we have had problems with TD-PSOLA.

REFERENCES

[1] E. B. George and M. J. T. Smith, "An analysis-by-synthesis approach to sinusoidal modeling applied to the analysis and synthesis of musical tones," *Journal of the Audio Engineering Society*, vol. 40, pp. 497–516, June 1992.

[2] E. B. George and M. J. T. Smith, "Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model," *IEEE Transactions on Speech and Audio Processing*, vol. 5, pp. 389–406, September 1997.

[3] S. G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictionaries," *IEEE Transactions on Signal Processing*, vol. 41, pp. 3397–3415, December 1993.

[4] T. F. Quatieri and R. J. McAulay, "Shape invariant time-scale and pitch modification of speech," *IEEE Transactions on Signal Processing*, vol. 40, pp. 497–510, March 1992.

[5] Y. Stylianou, J. Laroche, and E. Moulines, "High quality speech modification based on a harmonic + noise model," *Proceedings of EUROSPEECH*, pp. 451–454, September 1995.

[6] M. W. Macon and M. A. Clements, "Speech concatenation and synthesis using an overlap-add sinusoidal model," in *Proc. ICASSP*, vol. 1, pp. 361–364, May 1996.

[7] M. W. Macon, *Speech Synthesis Based on Sinusoidal Modeling*. PhD thesis, Georgia Institute of Technology, October 1996.

[8] M. W. Macon, L. Jensen-Link, J. Oliverio, M. A. Clements, and E. B. George, "A singing voice synthesis system based on sinusoidal modeling," in *Proc. ICASSP*, vol. 1, pp. 435–438, May 1997.

[9] A. W. Black and P. Taylor, "The Festival speech synthesis system: System documentation," Tech. Rep. HCRC/TR-83, Human Communication Research Centre, University of Edinburgh, Scotland, UK, January 1997. available from `www.cstr.ed.ac.uk/projects/festival.html`.

[10] E. B. George, "Practical high-quality speech and voice synthesis using fixed frame rate ABS/OLA sinusoidal modeling," in *Proc. ICASSP*, vol. I, pp. 301–304, May 1998.

[11] X. Serra and J. S. Smith, "Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition," *Computer Music Journal*, vol. 14, no. 4, pp. 12–23, 1990.

[12] M. W. Macon and M. A. Clements, "Sinusoidal modeling and modification of unvoiced speech," *IEEE Transactions on Speech and Audio Processing*, vol. 5, pp. 557–560, November 1997.

[13] T. F. Quatieri and R. J. McAulay, "Phase coherence in speech reconstruction for enhancement and coding applications," *Proc. ICASSP*, pp. 207–210, April 1989.

[14] C. Hamon, E. Moulines, and F. Charpentier, "A diphone synthesis system based on time-domain prosodic modificaitons of speech," in *Proc. of the International Conf. on Acoustics, Speech, and Signal Processing*, pp. 238–241, 1989.

[15] A. Syrdal, Y. Stylianou, L. Garrison, A. Conkie, and J. Schroeter, "TD-PSOLA versus Harmonic Plus Noise Model in diphone based speech synthesis," in *Proc. of the International Conf. on Acoustics, Speech, and Signal Processing*, vol. I, pp. 273–276, May 1998.