SPEECH CONCATENATION AND SYNTHESIS USING AN OVERLAP-ADD SINUSOIDAL MODEL

Michael W. Macon and Mark A. Clements

School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, GA 30332-0250
{*macon, clements*}*@ee.gatech.edu*

ABSTRACT

In this paper, an algorithm for the concatenation of speech signal segments taken from disjoint utterances is presented. The algorithm is based on the *Analysis-by-Synthesis/Overlap-Add* (ABS/OLA) sinusoidal model [1, 2, 3], which is capable of performing high quality pitch- and time-scale modification of both speech and music signals. With the incorporation of concatenation and smoothing techniques, the model is capable of smoothing the transitions between separately-analyzed speech segments by matching the time- and frequency-domain characteristics of the signals at their boundaries. The application of these techniques in a text-to-speech system based on concatenation of diphone sinusoidal models is also presented.

1. INTRODUCTION

One commonly used technique for synthesis of the speech waveform in text-to-speech synthesis is concatenation of short speech units taken from a prerecorded inventory. After concatenation, these units are modified in duration and "melody" to smoothly join each other and achieve the prosody of a natural utterance. In order to perform these modifications without introducing unnatural-sounding artifacts, signal modeling techniques, such as the popular PSOLA technique [4], must be employed. Sinusoidal signal models have been shown to be useful for speech prosody modification [3, 5] and speech synthesis [6], as well as music synthesis [2]. Specifically, the ABS/OLA sinusoidal model provides an attractive framework for speech concatenation and synthesis due to its computationally efficient overlap-add synthesis algorithm and its high quality speech modification capabilities. This paper describes the use of this model to concatenate and modify speech segments and the application of these techniques to speech synthesis.

In the ABS/OLA model, the input signal $s[n]$ is represented by a sum of overlapping short-time signal frames $s_k[n]$.

$$s[n] = \sigma[n] \sum_{k} w[n - kN_s] s_k[n] \tag{1}$$

where N_s is the frame length, $w[n]$ is a complementary window function that is nonzero over the interval $[-N_s, N_s]$, $\sigma[n]$ is a slowly time-varying gain envelope, and $s_k[n]$ represents the kth frame "synthetic contribution" to the synthesized signal. Each signal contribution $s_k[n]$ is represented as the sum of a small number of constant-frequency sinusoidal components, given by

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

where L is the number of sinusoidal components in the frame, and A_l^k, ω_l^k, and ϕ_l^k are the kth frame 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 [5].

Overlap-add synthesis is performed by a procedure that uses the inverse fast Fourier transform to compute each contribution $s_k[n]$, rather than sets of oscillator functions, as in [5]. Time-scale modification is achieved with the model by changing the time evolution rate of the model parameters for each frame $s_k[n]$ and changing the frame duration, while imposing a "quasi-harmonic" structure on the sinusoidal components to maintain general waveform shape characteristics. 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 speech formant structure is maintained [3].

This work was supported in part by Sanders Inc., a Lockheed-Martin Company.

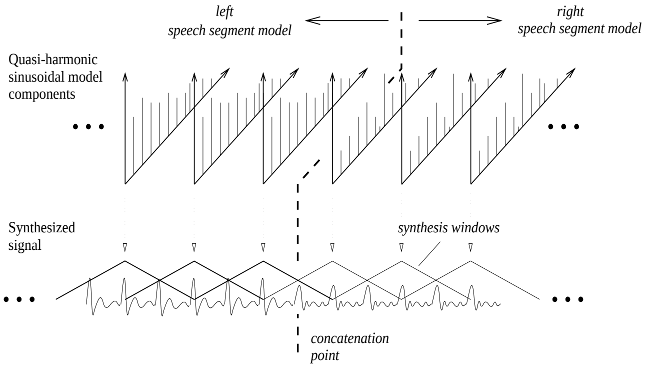

Figure 1: Concatenation of segments using sinusoidal model parameters

2. CONCATENATION OF MODELED SEGMENTS

The ABS/OLA sinusoidal model analysis generates two quantities that represent each input signal frame: (1) a set of quasi-harmonic sinusoidal parameters for each frame (with an implied fundamental frequency estimate), and (2) a slowly time-varying gain envelope. Disjoint modeled speech segments can be concatenated by simply stringing together these sets of model parameters and resynthesizing, as shown in Figure 1. Small waveform discontinuities at the concatenation point will be implicitly smoothed over by the nature of the overlap-add procedure. However, since the joined segments are analyzed from disjoint utterances, substantial variations between the time- or frequency-domain characteristics of the signals may occur at the boundaries. These differences manifest themselves in the sinusoidal model parameters. Thus, the goal of the algorithms described here is to make discontinuities at the concatenation points inaudible by directly manipulating the sinusoidal model components in the neighborhood of the boundaries.

In the problem presented here, it is assumed that the desired fundamental frequency contour for the utterance is provided *a priori*. In order to create a signal with the desired fundamental frequency contour, a pitch modification factor β is computed in each frame by $\beta = \omega_0^{desired}/\omega_0$, where ω_0 is a pitch estimate computed from each frame's sinusoidal components [1]. Applying this pitch shift to the sinusoidal model components introduces the desired pitch contour and implicitly matches the fundamental frequency across the boundary as well.

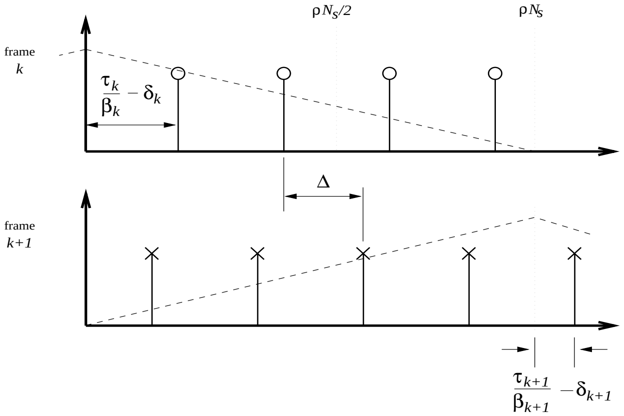

Figure 2: Pitch pulse alignment after time-scale and pitch modification

2.1. Pitch pulse alignment

Proper alignment of pitch pulses from frame to frame is essential to maintaining good resynthesized speech quality, and this alignment can be performed by adjusting the phases of the sinusoidal components prior to overlapping and adding successive frames. For connected-speech modification, a recursive equation for the necessary linear phase shift δ for a given frame is derived in [1, 3], based on the frame's time scale modification factor ρ, pitch modification factor β, and pitch period T_o, as well as a *pitch pulse onset time* [1] estimate τ. Implicit in the derivation of this equation is the assumption that successive pitch pulse onset times will be closely related to each other by the pitch period, but this assumption is not valid across the boundary of concatenated segments, since the segments come from disjoint utterances. Therefore, a new expression for δ is necessary.

Figure 2 shows schematically the relationships of pitch pulse contributions from two successive frames prior to overlap. To find an expression for the necessary time shift δ_{k+1} in this case, an expression is first found for the pitch pulse locations due to the contributions of frame k. These pulse locations will depend on the pitch pulse onset time for frame k, τ_k, the pitch modification factor β_k, the pitch period T_o^k, and the (previously computed) time shift that has already been applied to frame k, δ_k.[2] Similarly, the pulse locations due to frame $k+1$ can be found from τ_{k+1}, β_{k+1}, and T_o^{k+1}. Taking into account the modified frame duration ρN_s, the indices of the pitch pulses adjacent to

[1] The pitch pulse onset time [5] is the (hypothetical) location of the first pitch pulse in a given frame.

[2] Note that these are only hypothetical pulse locations – all analysis and synthesis is performed in the frequency domain.

the center of the overlap region, denoted $\hat{\imath}_k$ and $\hat{\imath}_{k+1}$, can be found.

The goal of the frame alignment process is to shift frame $k+1$ such that the pitch pulses of the two frames line up and the waveforms add coherently. A reasonable way to achieve this is to force the time difference Δ between the pitch pulses adjacent to the center to be the average of the modified pitch periods in the two frames. Typically, the modified pitch periods T_o^k/β_k and T_o^{k+1}/β_{k+1} will be approximately equal, since the purpose of concatenating segments is generally to produce natural sounding speech without discontinuities in the fundamental frequency contour. Thus the following expression can be derived

$$\delta_{k+1} = \delta_k + \frac{\tau_{k+1}}{\beta_{k+1}} - \frac{\tau_k}{\beta_k} + \hat{\imath}_{k+1}\left(\frac{T_o^{k+1}}{\beta_{k+1}}\right) - \hat{\imath}_k\left(\frac{T_o^k}{\beta_k}\right) + \rho N_s - \tilde{T}_o^{avg} \qquad (3)$$

This gives the necessary linear phase shift of the sinusoidal components in frame $k+1$ to give coherent overlap at the boundary. This time shift (which need not be an integer) can be implemented directly in the frequency domain by modifying the sinusoid phases ϕ_i prior to resynthesis.

The alignment algorithm described above has the desirable trait that all expensive computation (like finding τ_k) can be done off-line in the analysis phase. However, the concatenation results are quite sensitive to the accuracy of the pitch onset time estimate. The onset time estimation scheme described in [7] sometimes produces errant results, resulting in alignment errors that are clearly audible. To help overcome this problem, a post-processing step has been included in the sinusoidal analysis. This post-processor uses fundamental frequency estimates to find and correct gross errors in the onset time estimates.

2.2. Spectral smoothing

Another source of perceptible discontinuity across concatenation boundaries is mismatch in the signal spectral shape described by the sinusoidal amplitudes. It is assumed that the segments being joined are somewhat similar to each other in formant structure. However, differences in spectral content are often still present due to coarticulation and other effects.

In the ABS/OLA pitch modification algorithm, a spectral envelope estimate is used to maintain formant locations and spectral shape while frequencies of sinusoids in the model are altered. This envelope is computed from a set of cepstral features generated in the analysis process. An "excitation model" is computed by dividing each complex sinusoid amplitude by the spectral estimate $H(\omega)$ at the sinusoid frequency, ω. The excitation sinusoids are then shifted in frequency, and the spectral envelope is reintroduced to obtain the pitch-shifted signal. This operation also provides a mechanism for smoothing spectral differences over the concatenation boundary, since a *modified* envelope may be imposed on the sinusoidal components after pitch-shifting.

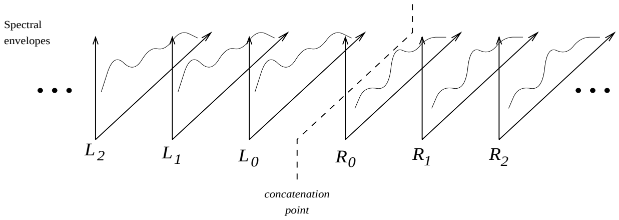

Figure 3: Cepstral envelope matching

Spectral differences across concatenation points are smoothed by adding weighted versions of the cepstral feature vectors from one side of the segment boundary to cepstral feature vectors from the other side, and vice-versa, to compute a new set of smoothed cepstral feature vectors, and a modified spectral envelope $H_k^s(\omega)$. Assume that cepstral features for the left-side segment, $\{..., \mathcal{L}_3, \mathcal{L}_2, \mathcal{L}_1, \mathcal{L}_0\}$ from right to left, and features for the right-side segment, $\{\mathcal{R}_0, \mathcal{R}_1, \mathcal{R}_2, \mathcal{R}_3, ...\}$, are to be concatenated as shown in Figure 3. Smoothed cepstral features $\mathcal{L}_k^s$ for the left segment and $\mathcal{R}_k^s$ for the right segment are found by

$$\begin{aligned} \mathcal{L}_k^s &= w_k \mathcal{L}_k + (1 - w_k)\mathcal{R}_0 \qquad (4) \\ \mathcal{R}_k^s &= w_k \mathcal{R}_k + (1 - w_k)\mathcal{L}_0 \end{aligned}$$

where

$$w_k = 0.5 + \frac{k}{2N_{smooth}}, \qquad k = 1, 2, ..., N_{smooth}$$

where N_{smooth} frames to the left and right of the boundary are incorporated into the smoothing. It can be shown that this linear interpolation of cepstral features is equivalent to linear interpolation of log spectral magnitudes. Prior to application of this smoothing algorithm, the cepstral features are normalized such that $\int H_k(\omega) d\omega = 1$ for all k, so that only spectral shape, not overall gain, is interpolated. After the "excitation" sinusoidal frequencies have been modified, each sinusoidal component in frame k is multiplied by $H_k^s(\omega)$ to impart the spectral shape derived from the smoothed cepstral features.

Comparisons of this smoothing procedure with unsmoothed concatenation indicate that it significantly reduces perceptual discontinuities between the joined segments.

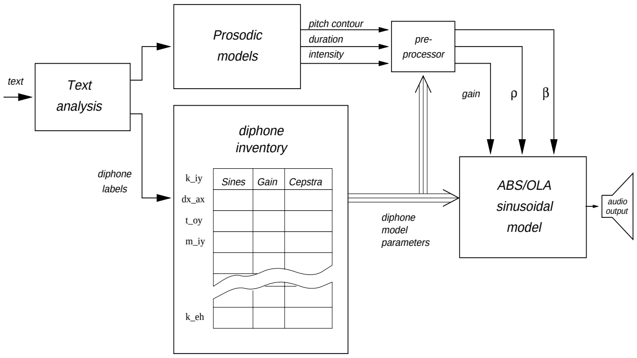

Figure 4: Text-to-speech synthesis by diphone concatenation

3. APPLICATIONS

The main application of these sinusoidal signal modification and concatenation algorithms has been as the waveform synthesis engine of a text-to-speech (TTS) system, as shown in Figure 4. In the system, a text analysis module is used to produce a list of diphone labels. This list is used to select units from an inventory of previously-analyzed diphone sinusoidal models, which represent the transitions between all possible phone pairs. The diphone units are collected by cutting segments from natural utterances containing each unit and submitting these to the sinusoidal analysis.

The sequence of diphone sinusoidal models is concatenated using the techniques described above. Durations of individual phones are controlled by compressing or expanding the time scale using the sinusoidal model. The "melody" of the utterance, output by an intonation model, is implemented by computing a frame-by-frame pitch modification factor which transforms the inventory unit pitch to the desired pitch. Likewise, the signal power is adjusted by modifying the sinusoid amplitudes.

On the analysis side, the model requires no hand marking of speech events such as pitch pulse locations. The system is capable of producing high-quality output speech. Furthermore, it provides a flexible *model* of the signal at its output, which can be useful for exploring other modifications such as glottal pulse shaping and speech style modification.

This framework also has application outside of diphone speech synthesis. For instance, concatenation of words or insertion of keywords into "canned" utterances[3] can be performed with natural prosodic contours imposed on the speech. Concatenation as an extension to the ABS/OLA sinusoidal music synthesis algorithm [2] for synthesis of musical instruments and singing voice is also under investigation.

4. SUMMARY

This paper has presented the application of the ABS/-OLA sinusoidal model to the concatenation of subword speech units taken from disjoint utterances. These techniques have been applied in a text-to-speech system based on concatenation of elements from an inventory of diphone transition units represented by the sinusoidal model. The model provides a flexible, efficient method for the concatenation and prosodic modification necessary in speech synthesis, and has many other practical applications in speech and music signal processing.

5. REFERENCES

[1] E. B. George, *An Analysis-by-Synthesis Approach to Sinusoidal Modeling Applied to Speech and Music Signal Processing.* PhD thesis, Georgia Institute of Technology, November 1991.

[2] 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.

[3] 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.* in review.

[4] E. Moulines and F. Charpentier, "Pitch synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication*, vol. 9, pp. 453–467, December 1990.

[5] 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.

[6] E. R. Banga and C. García-Mateo, "Shape-invariant pitch-synchronous text-to-speech conversion," *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 656–659, May 1995.

[7] T. F. Quatieri and R. J. McAulay, "Mixed-phase deconvolution of speech based on a sine-wave model," *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pp. 649–652, April 1987.

[3] for instance, in interactive voice response or announcement systems