

---

# Speech Synthesis Using Damped Sinusoids

**James M. Hillenbrand**

Department of Speech Pathology  
and Audiology  
Western Michigan University  
Kalamazoo

**Robert A. Houde**

RIT Research Corporation  
Rochester, NY

---

A speech synthesizer was developed that operates by summing exponentially damped sinusoids at frequencies and amplitudes corresponding to peaks derived from the spectrum envelope of the speech signal. The spectrum analysis begins with the calculation of a smoothed Fourier spectrum. A masking threshold is then computed for each frame as the running average of spectral amplitudes over an 800-Hz window. In a rough simulation of lateral suppression, the running average is then subtracted from the smoothed spectrum (with negative spectral values set to zero), producing a masked spectrum. The signal is resynthesized by summing exponentially damped sinusoids at frequencies corresponding to peaks in the masked spectra. If a periodicity measure indicates that a given analysis frame is voiced, the damped sinusoids are pulsed at a rate corresponding to the measured fundamental period. For unvoiced speech, the damped sinusoids are pulsed on and off at random intervals. A perceptual evaluation of speech produced by the damped sinusoid synthesizer showed excellent sentence intelligibility, excellent intelligibility for vowels in /hVd/ syllables, and fair intelligibility for consonants in CV nonsense syllables.

**KEY WORDS:** speech synthesis, spectral peaks, speech perception, vocoder

---

The last several decades have seen a proliferation of methods for the synthesis of high quality speech. Some of these techniques, such as the sinusoidal method introduced by McAuley and Quatieri (1986), produce speech that is of such high quality that it can be essentially indistinguishable from the original utterance upon which it is modeled. Despite these developments, formant synthesizers continue to be widely used in experiments that are aimed at shedding light on a variety of fundamental questions in speech perception. The continued reliance on formant synthesizers in speech research is due in part to the fact that the underlying control parameters—chiefly fundamental frequency, degree of periodicity, and formants—are widely assumed to have some level of psychological reality,<sup>1</sup> making the method suitable for exploring a wide range of issues in speech perception, such as context effects, cue trading, talker normalization, perceptual compensation for coarticulation, phonetic boundary effects, normalization for speaking rate, and a variety of related issues that have formed the core of speech perception research for several decades. Further, formant synthesizers are heavily used in studies investigating the neural representation of speech and in cochlear implant research, because they provide the ability to clearly and simply specify the ways in which various test stimuli differ from one another. Formant

---

<sup>1</sup>The precise psychological status of formant frequencies in human speech perception remains a matter of some debate, with some investigators arguing in favor of pattern matching based on the gross shape of the spectrum (e.g., Bladon, 1982; Bladon & Lindblom, 1981; Zahorian & Jagharghi, 1993; but see also Hedlin, 1982; Klatt, 1982).

synthesizers are not without their limitations, however. Though formant synthesis by rule of stylized words and syllables is often straightforward, copying complex naturally spoken utterances can be a cumbersome process because it is quite difficult to derive all of the relevant control parameters automatically from the original signal. Formant frequencies typically have to be measured or edited by hand (e.g., Hillenbrand & Nearey, 1999), and a considerable amount of trial and error may be needed to faithfully model certain speech features such as fricative spectra, nasal consonant murmurs, nasalized vowels, and stop consonant release bursts. The damped sinewave synthesizer (DSS) that is described in this paper shares certain features with a formant synthesizer but avoids many of the difficulties associated with deriving the relevant control parameters automatically from natural speech.

The speech analysis and resynthesis system that was developed operates by summing exponentially damped sinusoids with frequencies and amplitudes corresponding to peaks in the spectrum envelope. In principle, the technique is close to formant synthesis,<sup>2</sup> with two important exceptions: (1) the method does not require an explicit formant analysis (although spectral peaks are extracted that will, in many cases, correspond to formants), and (2) methods have been developed for deriving all relevant control parameters automatically from natural speech.

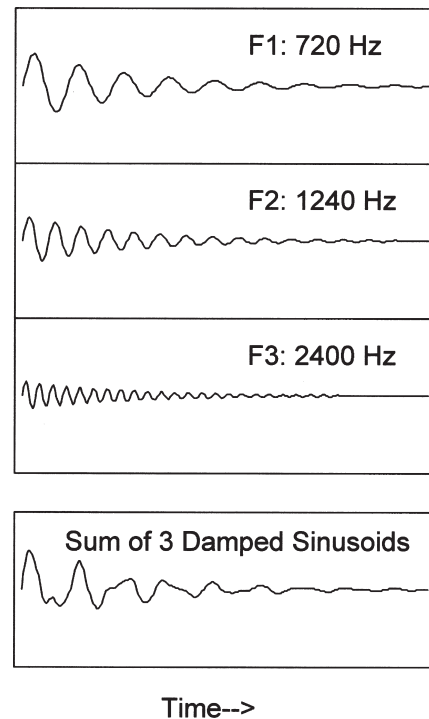
## Design Principles

In a conventional formant synthesizer, a simple sustained vowel is created by passing a glottal pulse train through series- or parallel-connected bandpass filters that simulate the effects of the formant resonators. Unvoiced speech is created by using some type of noise sequence to excite the formant resonators. The key principle underlying the damped sinewave synthesizer is that the same effect can be achieved directly in the time domain by summing exponentially damped sinusoids. Some aspects of damped sinusoid synthesis have been described in Witten (1982), Rodet (1984), and Hertrich and Ackermann (1999). In the time domain, each individual formant is an exponentially damped sinusoid of the form:

$$d(t) = a e^{-bt} \sin(2\pi ft) \text{ (for } t \geq 0),$$

where  $a$  = amplitude,  $t$  = time,  $f$  = frequency (Hz), and  $b$  = bandwidth (Hz). Figure 1 shows a single cycle of the vowel /a/ generated by summing damped sinusoids at frequencies corresponding to  $F_1$ ,  $F_2$ , and  $F_3$ . The bandwidth of

**Figure 1.** Formants in the time domain: Creating a single cycle of the vowel /a/ by summing exponentially damped sinusoids at frequencies corresponding to  $F_1$ ,  $F_2$ , and  $F_3$ .



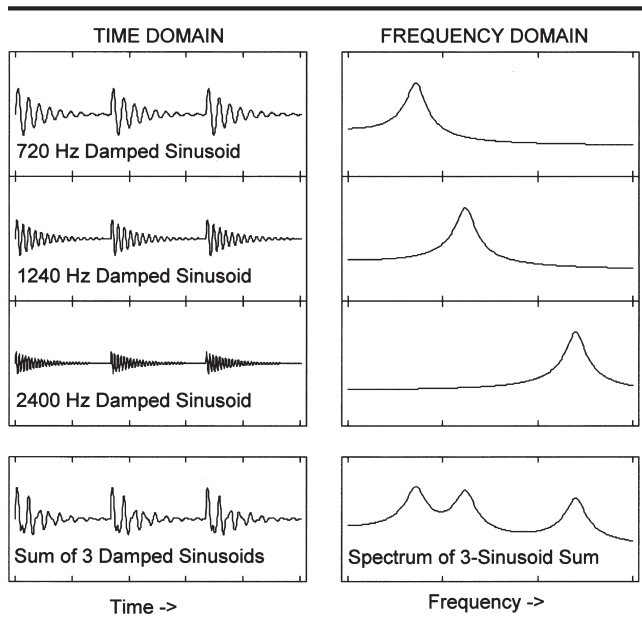
each formant is controlled by varying the time constant of the damping function; short decay times produce broad bandwidths and long decay times produce narrow bandwidths. Formant amplitude relationships are controlled simply by scaling the amplitudes of the individual sinusoids before summing.

Figure 2 illustrates the summing operation for a short segment of a sustained /a/. The time-domain summing operation that is shown on the left is equivalent to the frequency-domain operations that are shown to the right. Voiced speech is produced by pulsing the damped sinusoids at onset-to-onset intervals corresponding to the fundamental period. It is important to note that, except in the case of very long periods (or very short damping constants), adjacent damped oscillations will overlap; that is, a given damped oscillation will typically ring into the following cycle. The summing method that is used is a simple overlap-and-add of the kind that is used in techniques such as PSOLA (Moulines & Charpentier, 1990). The summing operation illustrated in Figure 2 is, of course, equivalent to convolving a sequence of impulses, spaced at the fundamental period, with an impulse response consisting of the sum of damped sinusoids at spectral peak frequencies.

Witten's (1982) description of damped sinusoid synthesis explicitly states that the method is suitable only

<sup>2</sup>We have chosen to describe the damped sinewave synthesizer as a variation of formant synthesis, but this is largely a matter of expository convenience. As we will discuss later, the damped sinewave synthesizer also has a good deal in common with the spectral envelope estimate vocoder developed by Paul (1981), the homomorphic vocoder described by Oppenheim (1969), and several other source-filter-based channel vocoders.

**Figure 2.** Synthesizing a short segment of a sustained /a/ by summing exponentially damped sinusoids at frequencies corresponding to  $F_1$ ,  $F_2$ , and  $F_3$  and by pulsing the damped sinusoids at onset-to-onset intervals corresponding to the fundamental period. Note that the time-domain summing operations shown to the left are equivalent to the frequency-domain operations that are shown to the right.



for voiced speech. Natural-sounding aperiodic speech can, in fact, be produced with the damped sinusoid summing method simply by replacing the periodic summing interval that is illustrated in Figure 2 with a random interval. In our implementation, this is accomplished by convolving the summed, damped sinusoids with a source signal consisting of a sequence of impulses, generated at the sample period, whose amplitudes are either zero or nonzero with a probability of .5 (the control of amplitude for the pulse sequence will be described in detail below). Conceptually, this method is similar to the technique used by early Haskins researchers to simulate aperiodic sounds on the Pattern Playback (Cooper, Delattre, Liberman, Borst, & Gerstman, 1952), whose sound source consisted exclusively of a 120-Hz periodic pulse train. By painting a stippled pattern on regions of the acetate belt where aperiodic segments were desired, harmonics of the source signal were pulsed on and off at quasi-random intervals, resulting in an aperiodic signal whose spectral shape was controlled by the frequency locations of the stippling. The damped sinusoid method can also be used to synthesize sounds such as voiced fricatives and breathy vowels that consist of both periodic and aperiodic components. This is done simply by mixing periodically generated damped sinusoids with those generated at random intervals at whatever voiced/unvoiced mixing ratio is desired.

## Copy Synthesis Using the Damped Sinewave Synthesizer

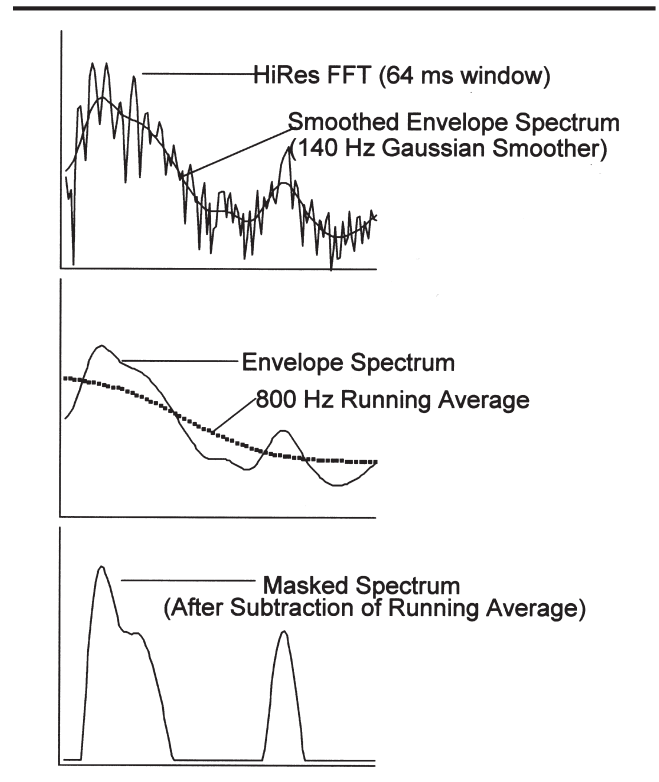
Control parameters for the damped sinewave synthesizer might be set by hand, as they are in the Klatt formant synthesis programs (Klatt, 1980; Klatt & Klatt, 1990); they might be derived by rule, as they are in text-to-speech systems (e.g., Klatt, 1987); or they might be derived automatically based on an analysis of the acoustic properties of a naturally spoken utterance. This paper will focus on the derivation of control parameters through acoustic analysis. The control parameters that are needed to drive the damped sinewave synthesizer consist of: (a) frequencies, amplitudes, and bandwidths of peaks in the spectrum envelope, (b) instantaneous fundamental frequency, (c) overall signal amplitude, and (d) degree of signal periodicity (i.e., voiced/unvoiced ratio). All four types of parameters are in widespread use in many areas of speech processing, and there is a quite large array of methods that might be used in a nearly infinite variety of combinations to derive these parameters. There are, for example, quite literally hundreds of methods that are available for measuring fundamental frequency and the degree of periodicity, as well as a wide range of choices for deriving the spectrum envelope (e.g., linear predictive coding, cepstral smoothing, cosine basis vectors, etc.) and for extracting peaks from the envelope. We attach no particular importance to the necessarily very specific set of choices that we adopted for the demonstration that is described here. Along with the signal processing methods themselves, many decisions had to be made regarding parameters such as the spectrum-analysis window size, the widths of various smoothing windows, update intervals, thresholds of various types, and so on. Settings for these kinds of parameters were determined through trial-and-error testing, comparison of original and resynthesized utterances, and a great deal of informal listening. In no sense are either the signal processing methods or the specific parameter values optimized, and we do not mean to attach any great importance to the detailed choices that we made regarding these settings. What we intend to show with the copy synthesis demonstration described below is that it is possible in principle to produce intelligible speech through a process in which the necessary DSS control parameters are derived automatically from naturally spoken utterances.

### Extraction of Envelope Peaks

Spectrum analysis begins with the calculation of a high-resolution Fourier spectrum (with no high-frequency pre-emphasis) over a 64-ms Hamming window

(512 points at the 8-kHz sample rate that was used to generate all of the examples that are discussed here). The window is advanced 10 ms each analysis frame. As shown in Figure 3, an estimate of the spectrum envelope is then derived by smoothing the narrowband spectrum with three passes of a 140-Hz running averager (an approximation to a Gaussian-weighted smoothing window).<sup>3</sup> In most cases, spectral peaks can be measured directly from the envelope spectrum. We have adopted a slightly more elaborate method to handle special cases such as the one illustrated in Figure 3. The spectra shown in this figure were taken from the vowel /o/ spoken by a man, with an expected second formant at about 900 Hz. Note, however, that the soft shoulder corresponding to  $F_2$  is not a spectral peak. To enhance contrast for spectra such as these, a masking threshold is computed as three passes of an 800-Hz running average of the envelope spectrum (middle panel of Figure 3). The 800-Hz running average is then subtracted from the envelope spectrum, with all values below the masking threshold set to zero.<sup>4</sup> Note that the soft shoulder corresponding to  $F_2$  emerges as a peak in the masked spectrum. A further benefit of the masking operation is that minor envelope peaks that fall below the masking threshold are rejected (e.g., note the minor envelope peak just below  $F_3$  in the middle panel of Figure 3, which lies below the masking threshold and therefore does not appear in the masked spectrum). Spectral peak frequencies are then measured from the masked spectrum in a straightforward way. No limit is placed on the number of peaks that are extracted in each frame. The number of peaks

**Figure 3.** Signal processing steps used to create masked spectra (up to 3 kHz only) for one frame of the vowel /o/. Top panel: High-resolution (64 ms) Fourier spectrum and smoothed envelope. Middle panel: Smoothed envelope and heavily smoothed running average, which serves as the masking threshold. Bottom panel: Masked spectrum, obtained by subtracting the masking threshold from the envelope spectrum and setting negative spectral values to zero.



<sup>3</sup> Here and elsewhere, “Gaussian-weighted running average” refers to an approximation implemented with three passes of a rectangular (i.e., unweighted) running average. In this smoothing operation, each spectral amplitude is replaced by the weighted average of  $n$  neighbors of higher and lower frequency, with the  $n$  being determined by the width of the smoothing window. Greater weight is assigned to spectral values at the center of the averaging window than to values nearer to the edge of the window. In a true Gaussian-weighted average, the distribution of weights follows a Gaussian function. A simple-to-implement close approximation to a Gaussian-weighted average can be achieved by running three passes of a rectangular average; that is, the output of an initial running-average operation becomes the input to a second running average, whose output in turn becomes the input to a third running average. A simple end-correction scheme is used in which the averaging window size is initially set to one point at either the left or right edge and the window size is successively expanded until the running average has shifted far enough so that  $n$  points are available.

<sup>4</sup> The type of simultaneous masking that is being (roughly) simulated here is based on the well-known tendency of highly active neurons to inhibit or suppress less active neurons in neighboring regions. The degree to which region  $a$  will mask region  $b$  depends on: (1) the average level of activity in region  $a$  and (2) the distance between the two regions. Both of these features are captured by a center-weighted average (implemented here with a Gaussian running average); that is, it clearly reflects average amplitude, but the masking influence exerted by a particular frequency region falls off systematically with increasing distance. For the specific purposes to which the masking operation is being applied here, the important point is that the operation increases contrast and makes it more likely that the peak picker will not miss perceptually relevant soft shoulders such as the one illustrated in Figure 3.

per frame typically averages about 5, with the maximum number of peaks per frame in a given utterance seldom exceeding about 12. No attempt is made to determine whether peaks do or do not correspond to formants and no continuity constraints are imposed on the evolution of spectral peaks over time. Spectral peak amplitudes are derived from the envelope spectrum rather than the masked spectrum. In our current implementation, bandwidths are not measured at all. We have found little perceptual advantage to controlling peak bandwidths, so they have been fixed at a constant 80 Hz for all peaks.

The processing stages described above are illustrated in the sequence of spectrograms shown in Figure 4. Note that the formant structure of the utterance is preserved reasonably well in the spectral peak sequence shown in the bottom panel, but there are also many peaks that clearly do not correspond to formants.

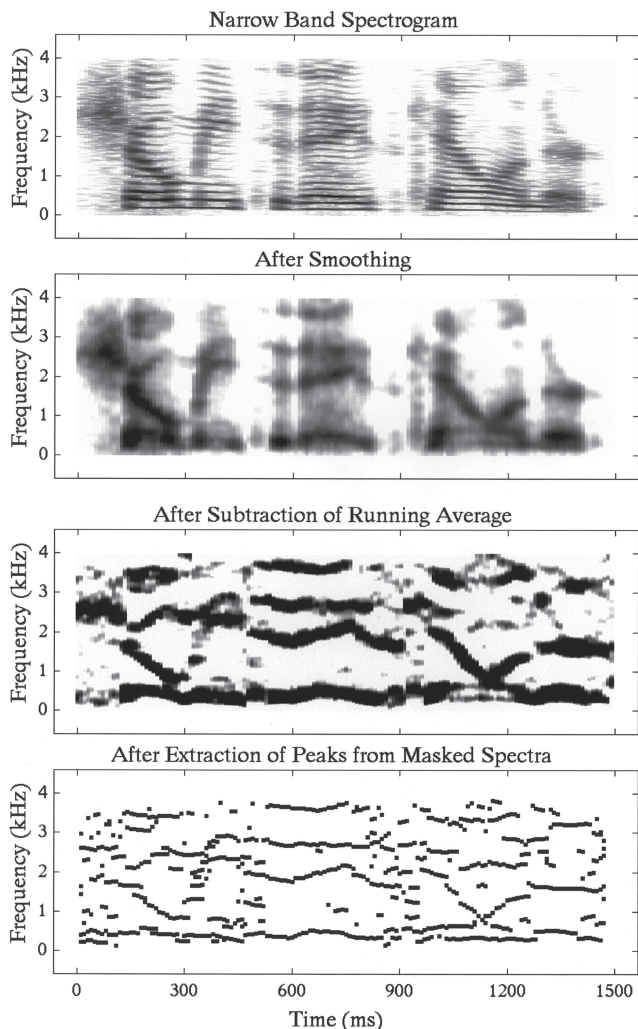
The end result of the spectral-peak analysis is the creation of a data structure such as the one shown in Figure 5. Each speech frame holds the sum of the damped oscillations for all the envelope peaks in that frame. The

damped sinusoid sums are 32 ms in duration and are updated with every 10-ms frame. The sequence of summed damped oscillations constitutes the time-varying impulse response of the vocal tract filter (or, more accurately, the combined effects of vocal tract filtering and the shape of the source signal) estimated entirely from spectral peaks. Each sum is scaled to a constant amplitude. As will be discussed below, scaling for overall amplitude occurs when the source signal is generated.

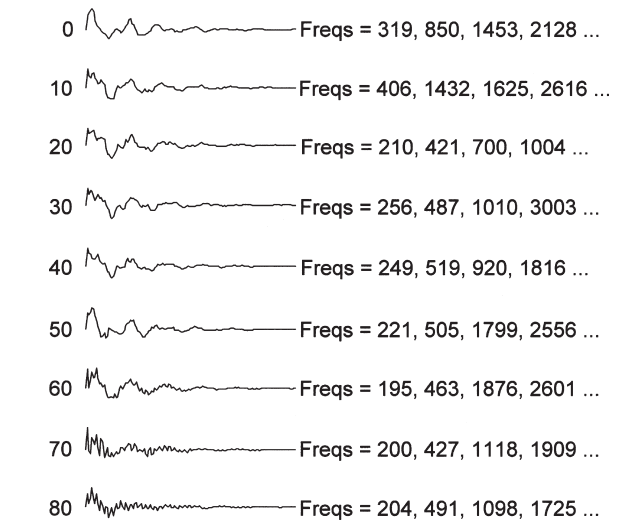
## Measurement of Fundamental Frequency and Periodicity

There is, of course, an exceptionally large and diverse set of methods available for estimating  $F_0$  and periodicity (for reviews, see Hess, 1983; Indefrey, Hess, & Seeser, 1985; O’Shaughnessy, 2000). The method we

**Figure 4.** Signal processing steps, beginning with a narrowband spectrum (top panel) and ending with the extraction of spectral peaks (bottom panel).



**Figure 5.** A small portion of the time-varying impulse response of the finite-impulse-response filter that is created by summing damped sinusoids at frequencies corresponding to peaks in the masked spectra.

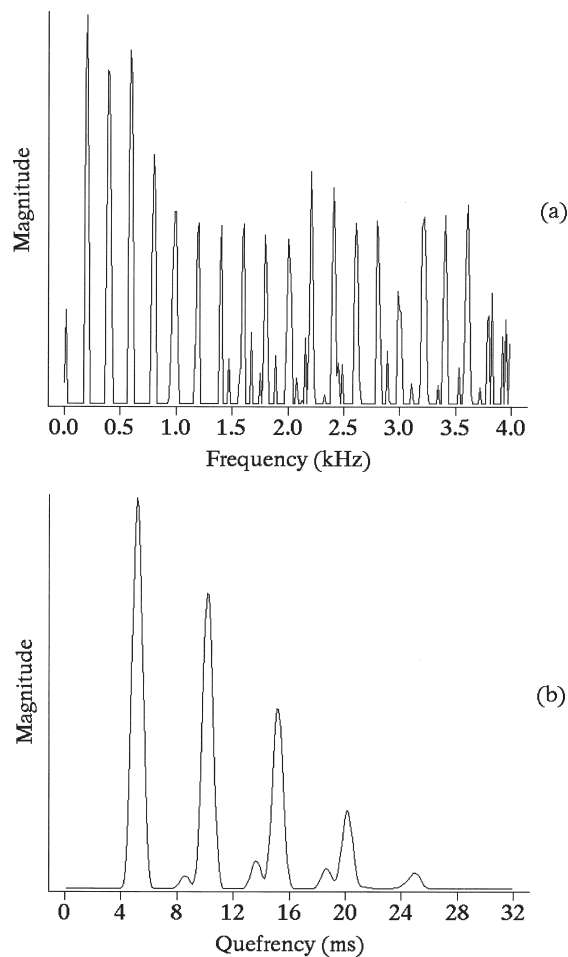


have adopted is a double-transform technique based on a variation of the cepstrum. As shown in the top panel of Figure 6, the first transform is a 512-point (64 ms) log amplitude spectrum. A 172-Hz Gaussian-weighted running average is then subtracted from the log spectrum, with negative values set to zero. This masking operation has the effect of flattening the spectrum and eliminating some minor spectral peaks. A cosine transform<sup>5</sup> is then computed from the masked log spectrum, and the output is halfwave rectified and lightly smoothed with a 5-bin (6.25 ms) Gaussian-weighted running average (see bottom panel of Figure 6). The fundamental period is defined as the period corresponding to the largest peak in the second transform, with the search constrained to the range 2.6–25 ms (40–381 Hz).  $F_0$  is measured every speech frame, regardless of the degree of signal periodicity; continuity constraints are not used; and there is no smoothing of  $F_0$  values over time.

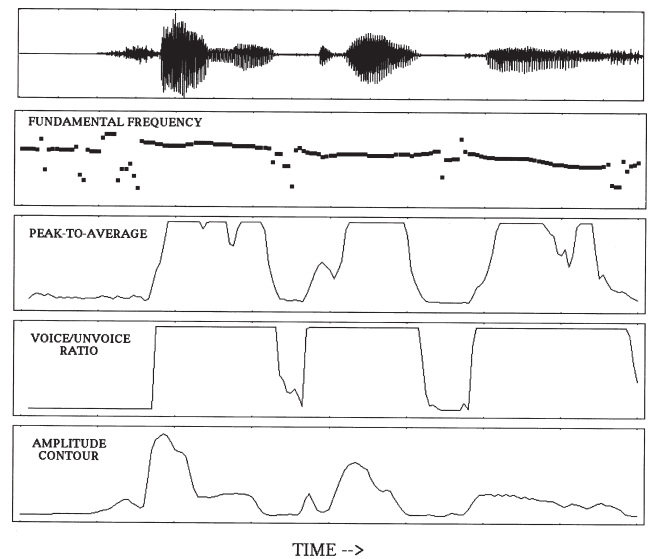
A peak-to-average ratio is calculated as the amplitude of the peak corresponding to the fundamental period divided by the average amplitude. A periodicity function, which controls the voiced/unvoiced mixing ratio, is derived from the peak-to-average ratio by a nonlinear mapping in which (a) values below 0.1 are set to 0.0, (b) values above 0.2 are set to 1.0, and (c) values between 0.1 and 0.2 are linearly scaled between 0.0 and 1.0.

<sup>5</sup>Note that the second transform is a cosine transform instead of the magnitude of the real and imaginary parts of the complex FFT, as is used in the conventional cepstrum. The reason is that we wish to decompose the spectrum into a set of sinusoidal functions whose maxima occur at harmonic frequencies in that spectrum, and cosine functions do just that. Hence the cosine function that best fits the spectrum corresponds to the largest periodic component in the spectrum.

**Figure 6.** Top panel: The first transform of a double-transform  $F_0$  and periodicity scheme, a 64-ms log Fourier spectrum after the subtraction of a 172-Hz Gaussian-weighted running average, with negative spectral values set to zero. Bottom panel: The second transform, a cosine transform of the spectrum shown in the top panel.



**Figure 7.** Illustration of some of the signal processing steps used to generate the source waveform for the utterance, "Show me today's calendar."

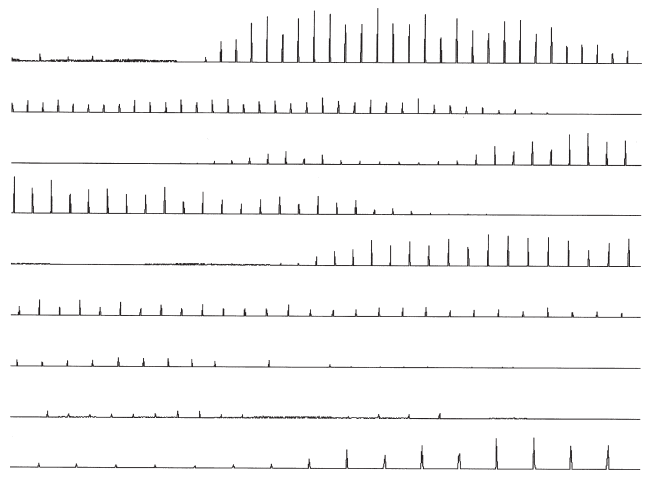


function and the aperiodic pulse sequence is amplitude modulated by its complement. The periodic and aperiodic waveforms are then mixed and the sum is amplitude modulated by the amplitude contour measured from the original signal. The amplitude contour is computed as the Gaussian-weighted running average of the fullwave rectified time signal, using a 20-ms averaging window. Approximately the first half (900 ms) of the source waveform for the utterance "Show me today's calendar" is shown in Figure 8.

## Generation of the Source Waveform

A source signal consisting of a sequence of single-sample discrete pulses varying in amplitude is created from the  $F_0$ , voiced/unvoiced mixing ratio, and amplitude functions (see Figure 7). The periodic and aperiodic components of the source signal are generated separately and then mixed. Initially, the periodic component consists of a sequence of constant-amplitude pulses spaced at the fundamental period, whereas the aperiodic component consists of a sequence of constant-amplitude pulses spaced at random intervals, with a probability of a nonzero pulse set to 0.5 at each sample point. Before any other scaling, the ratio of the peak amplitude of the periodic pulses to that of the aperiodic pulses is 0.3, resulting in roughly equal rms levels for the two pulse sequences. The periodic pulse sequence is then amplitude modulated by the voiced/unvoiced mixing

**Figure 8.** The first 900 ms of the source waveform for the utterance, "Show me today's calendar." Time is advancing from left to right and from top to bottom.



## Source-Filter Synthesis

Synthesis is carried out by a general-purpose program that convolves a source signal such as that shown in Figure 8 with a time-varying impulse response such as the one shown in Figure 5. Spectrograms of original and resynthesized versions of the utterance “Show me today’s calendar” are shown in Figure 9. There are, of course, many differences in spectral detail between the original and resynthesized utterances, but the two versions match reasonably well in the formant regions.

## Intelligibility of Speech Produced by the Damped Sinewave Synthesizer

A preliminary investigation was carried out to measure the intelligibility of speech produced by the DSS. Three separate participant groups were tested in tasks aimed at measuring vowel, consonant, and sentence intelligibility. In all three tests, listeners were graduate or undergraduate students in speech-language pathology and audiology.

### Speech Material

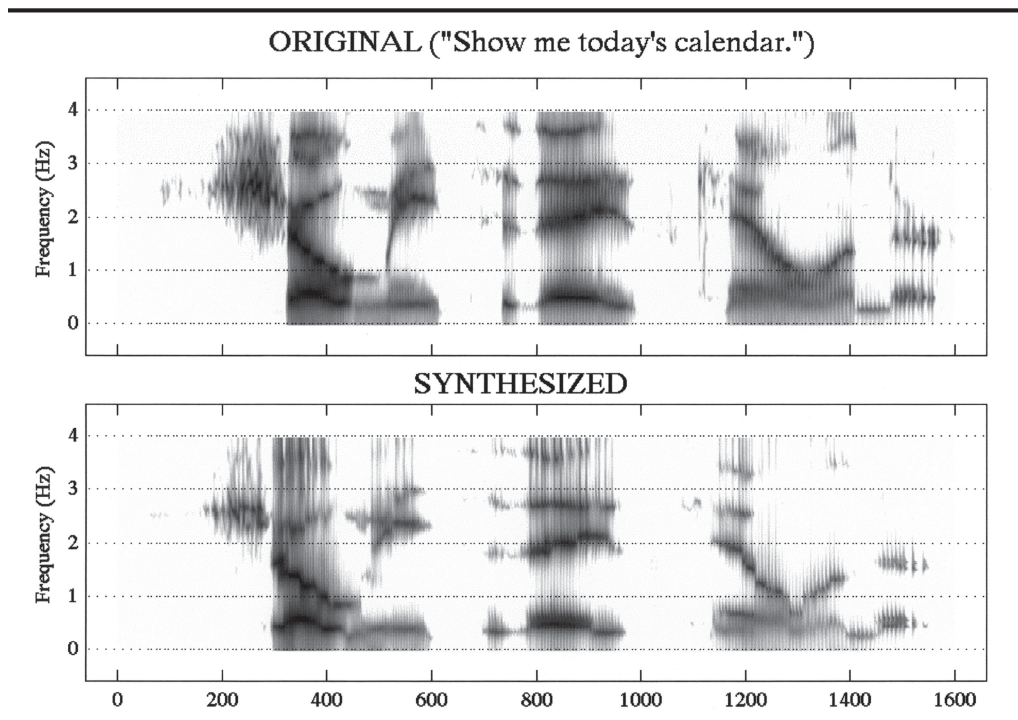
1. *Vowel Database.* The test signals for the vowel intelligibility test consisted of 300 /hVd/ utterances that were sampled from the 1,668 utterances recorded by Hillenbrand et al. (1995). The full database consists of

recordings of 12 vowels (/i, ɪ, e, ε, æ, a, ɔ, o, u, ʌ, ɜ, ɝ/) in /hVd/ syllables spoken by 45 men, 48 women, and 46 10- to 12-year-old children. The 300-stimulus subset that was selected from the larger database consisted of 25 tokens of each of the 12 vowels, with at least 1 token from 123 of the 139 talkers, and roughly equal numbers of tokens spoken by men, women, and children (see Hillenbrand & Nearey, 1999, for additional details). The test signals, which were originally digitized at 16 kHz, were digitally lowpass filtered at 3.8 kHz and down-sampled to 8 kHz.

2. *Consonant Database.* The test signals for the consonant intelligibility test consisted of 276 CV syllables sampled from the consonant recognition database described by Shannon, Jansvold, Padilla, Robert, and Wang (1999). The subset of the Shannon et al. database used in the present study consisted of CV syllables formed by 23 consonants (/b, d, g, p, t, k, m, n, l, r, f, v, θ, ð, s, z, ʃ, ʒ, ʒ, j, w, h/) and 3 vowels (/a, i, u/) spoken by 2 men and 2 women. The test signals, which were originally recorded at 44.1 kHz, were digitally lowpass filtered at 3.8 kHz and down-sampled to 8 kHz.

3. *Sentence Material.* Participants were tested on sentence-length utterances drawn from two sources. One source consisted of the 250 sentences that make up the HINT (Hearing In Noise Test) database described by Nilsson, Soli, and Sullivan (1994). These utterances were carefully spoken by a single male talker. A second sentence test, which was expected to be more difficult,

**Figure 9.** Gray-scale spectrograms of original and DSS synthesized versions of the utterance, “Show me today’s calendar.”



consisted of 50 utterances drawn from the TIMIT (Texas Instruments Institute of Technology) continuous speech database. The 50 sentences were drawn at random from the “phonetically diverse” subset of the larger database and included sentences spoken by 25 men and 25 women. The 16-kHz utterances were lowpass filtered at 3.8 kHz and down-sampled to 8 kHz.

## Procedures

1. *Vowel Test.* Test signals were presented for identification to 10 students in the speech-language pathology program at Western Michigan University. Each listener was asked to identify three versions of each of the 300 /hVd/ utterances, all sampled (or resampled) at 8 kHz: (1) the naturally produced utterance, (2) the synthetic version generated with the DSS, and (3) a formant-synthesized version generated with the Klatt and Klatt (1990) formant synthesizer. The formant-synthesized utterances were generated from hand-edited fundamental frequency and formant tracks measured in Hillenbrand, Getty, Clark, and Wheeler (1995), along with some simple rules for generating generic initial /h/ and final /d/ segments (see Hillenbrand & Nearey, 1999, for details). The order of presentation of the 900 signals was fully randomized (i.e., not blocked by talker or stimulus type), and the presentation order was shuffled separately for each listener. Participants were tested individually in a quiet room in a single session lasting about 50 minutes. Stimuli were scaled to maximum peak amplitude, lowpass filtered at 3.8 kHz at the output of a 16-bit D/A converter, amplified, and delivered to a single loudspeaker (Boston Acoustics A60) positioned about 1 m from the listener’s head at an average intensity of approximately 77 dBA. Listeners responded by pressing 1 of 12 keys on a computer keyboard that had been labeled with phonetic symbols for the vowels. The listening test was self-paced, and participants could repeat a stimulus as many times as they wished before entering a response. Each listening test was preceded by a brief practice session to ensure that listeners understood the task and interpreted the phonetic symbols appropriately.

2. *Consonant Test.* Listeners consisted of 11 speech-language pathology students. Participants identified both natural and synthetic versions of each of the 276 test signals (23 consonants  $\times$  3 vowels  $\times$  4 talkers) presented in random order and scrambled separately for each participant. Listeners responded by pressing 1 of 23 keys on an appropriately labeled computer keyboard. Other aspects of the procedure and instrumentation were identical to those described for the vowel test above.

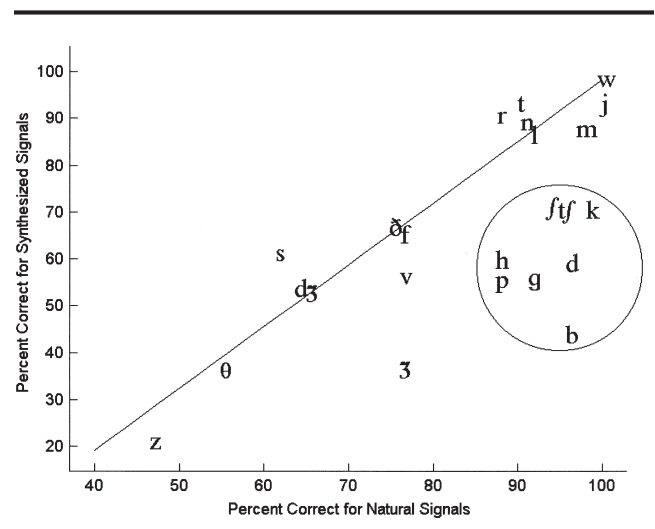
3. *Sentence Tests.* Listeners consisted of 16 speech-language pathology students who were tested on both

the HINT and TIMIT sentences, with the order of presentation counterbalanced across participants. For the HINT sentences, participants heard each of the 250 utterances in random order and were asked to repeat the sentence aloud to an experimenter who was seated next to the participant. On any given trial, the participant might hear the naturally produced utterance or the damped sinewave synthetic version, with a 50-50 probability. Participants were given the option of hearing each utterance a second time before repeating it back. Scoring was done online by the experimenter on a word-for-word basis. The presentation order of the 250 sentences and the ordering of synthetic versus natural versions of the sentences were randomized separately for each listener. The same basic procedure was used for the 50 TIMIT sentences.

## Results

1. *Vowels.* Vowel intelligibility averaged across all listeners and all vowels was 95.3% ( $SD = 4.7$ ) for the natural signals, 92.6% ( $SD = 4.9$ ) for the damped sinewave synthesized signals, and 90.0% ( $SD = 6.6$ ) for the formant synthesized signals. A one-way repeated measures ANOVA on arcsine-transformed percent correct scores showed a highly significant effect for stimulus type,  $F(2, 8) = 36.0, p < .001$ . Newman-Keuls post hoc tests showed that all pairs of conditions were reliably different, including the very slightly higher intelligibility of the damped sinewave stimuli compared with the formant synthesized versions of the same signals. Confusion matrices for all three sets of signals were unremarkable and quite similar to one another in overall form. For example, vowel-by-vowel percent correct values (i.e., the main diagonal in the confusion matrix) for the damped sinewave

**Figure 10.** Scatter plot of average identification rates for original and synthesized versions of the 23 consonants that were tested.



stimuli correlated at .87 with corresponding values for the natural signals, and cell-by-cell percentages for the damped sinewave stimuli correlated at .99 with corresponding values for the natural signals.

2. *Consonants.* Average intelligibility scores for the consonants were 84.3% ( $SD = 3.3$ ) for the natural signals and 66.4% for the damped sinewave stimuli. This difference is highly significant by a correlated  $t$  test ( $t = 23.7$ ,  $df = 10$ ,  $p < .001$ ). A complete analysis of the consonant confusion data is beyond the scope of this study, but some insight into the nature of the consonant transmission loss that is involved in the DSS can be gained by examining the average percent correct scores by consonant for the natural versus synthetic signals. Figure 10 shows a scatter plot of percent correct values for the natural and synthetic signals for each of the 23 consonants. There appear to be two reasonably distinct clusters of consonants. One consists of the cluster of data points near the solid line, which is a linear regression line fit to the data points for /m, n, l, r, f, v, θ, ð, s, z, ʒ, j, w, t/. The transmission loss for this cluster of consonants is relatively modest, ranging from a few percent in the upper right to some 20% in the lower left, and averaging about 8% overall. We speculate that this is due to an overall loss in spectral detail that results from estimating the shape of the spectral envelope entirely from peaks. Lying below the regression line is a cluster of consonants, including /b, d, g, p, k, ʃ, ʒ, h/, for which the transmission loss is greater than that predicted by the intelligibility relationship for the remaining consonants. This cluster includes the affricate /tʃ/ as well as all of the stop consonants except /t/. Our guess is that the transmission loss for these sounds is in part a side effect of the rather large 64-ms analysis window size, which might not do an adequate job of capturing the rapid spectral changes associated with these sounds. There is clearly more to the story than this, however, because (a) the cluster below the regression line includes /h/, /ʒ/, and /ʃ/, which ought to be captured reasonably well by a large analysis window; and (b) the main cluster near the regression line includes rapid spectrum-change sounds such as /t/ and /dʒ/.

An additional point worth noting about Figure 10 is the relatively high error rate for a few of the naturally spoken consonants, especially the high frequency fricatives and affricates /s, z, θ, ʒ/. This is undoubtedly a side effect of the 8-kHz sample rate, restricting the signal bandwidth to ~4 kHz.

3. *Sentences.* Both the naturally spoken ( $M = 99.9\%$ ,  $SD = 0.09$ ) and synthesized ( $M = 98.9\%$ ,  $SD = 0.6$ ) HINT sentences were recognized almost perfectly. The small drop in intelligibility for the synthesized utterances was nonetheless significant ( $t = 11.2$ ,  $df = 15$ ,  $p < .0001$ ). Similarly, both the natural ( $M = 98.2\%$ ,  $SD = 1.6$ ) and synthetic ( $M = 94.2\%$ ,  $SD = 3.4$ ) versions of the TIMIT

sentences were highly intelligible, and the difference between the two recognition rates was significant ( $t = 7.1$ ,  $df = 15$ ,  $p < .0001$ ).

## Discussion

A speech synthesizer was developed that is based on the reconstruction of speech by the summation of exponentially damped sinusoids at frequencies and amplitudes corresponding to peaks in the spectrum envelope. The damped sinusoids are pulsed at onset-to-onset intervals corresponding to the fundamental period during voiced intervals and at random intervals during unvoiced speech segments. Periodic and randomly spaced pulses can also be combined at variable mixing ratios to simulate speech signals with mixed periodic/asperiodic sources. In the implementation described here, envelope peaks were extracted from smoothed, masked Fourier spectra, whereas instantaneous fundamental frequency and voiced/unvoiced mixing ratio were estimated using a double-transform method similar to the cepstrum. A preliminary perceptual evaluation of the DSS showed excellent vowel intelligibility, fair consonant intelligibility, and excellent intelligibility of both easy sentences and sentences of moderate difficulty.

Given the considerable familiarity of the speech research community with formant synthesizers, we chose to introduce the damped sinewave synthesis method as a variation on formant synthesis. There are, of course, some important differences between damped sinewave synthesis and formant synthesis, both in practical and conceptual terms, and a rather different set of advantages and drawbacks accrue to each of these approaches to synthesis. As noted earlier, an important advantage of the damped sinewave method lies in the ability to drive the synthesizer with unedited spectral peaks that may or may not correspond to formants. From a practical point of view, this means that the spectral peaks that drive the DSS do not need to show the slot-by-slot continuity that is required by a formant synthesizer. For example, in the bottom panel of Figure 4, note that there is an abrupt change in the spectral peak structure of the utterance "Show me today's calendar" toward the end of the vowel /e/ of "today." At about 700 ms, there is a sequence of about five frames in which a spectral peak suddenly emerges between  $F_1$  and  $F_2$  (i.e., at about 1000 Hz), then abruptly disappears some 50 ms later. (Many other sudden changes in spectral peak structure can be observed in Figure 4, and complexities such as these lie at the heart of the still-unresolved formant tracking problem.) Discontinuities such as these in formant synthesis control parameters produce severe distortions in the synthesized speech signal as a result of the abrupt changes that occur across frame boundaries in the coefficients of the digital formant resonators. The DSS gets

around this kind of problem by avoiding the formant synthesizer's requirement for assigning spectral peaks to specific formant numbers, effectively using the spectral peaks that are extracted from a given frame to estimate the gross shape of the spectral envelope for that frame. When spectral peaks abruptly come and go, as in the example from Figure 4, the filter coefficients do, in fact, change across frame boundaries; however, the changes are relatively modest and do not result in the severe signal distortions that would occur in a formant synthesizer. When the filter component of the DSS method is thought of in terms of spectral envelope estimation, our synthesizer has a good deal in common with methods such as the spectral envelope estimation vocoder described by Paul (1981), the homomorphic vocoder described by Oppenheim (1969), and a large family of source-filter-based channel vocoders going back at least as far as Dudley (1939). The primary difference between our approach and earlier work with channel vocoders is that the present method produces estimates of the spectral envelope shape entirely from the relatively small number of smoothed spectral peaks (typically on the order of about 5 peaks per frame, on average), producing a reasonably good match between the original and reconstructed spectral envelope in the peak regions but allowing for only a rough match in nonpeak regions, which are considerably less relevant to phonetic quality (e.g., Hedlin, 1982; Klatt, 1982).

It is worth noting that, although the DSS does not require the kind of formant continuity that is demanded by a formant synthesizer, the damped sinewave synthesizer can, in fact, be driven by formant contours in the traditional sense. As with formant synthesis, the formant parameters can be measured from a naturally spoken utterance, they can be generated by rule, or a hybrid approach can be used in which formant parameters that are measured from a naturally spoken utterance are modified in specific ways in order to test hypotheses about the cues underlying phenomena such as phonetic recognition, speaker identity, or perceived prosodic characteristics. In our view, the damped sinewave approach offers some important advantages over conventional formant synthesis if the hybrid method is used because the derivation of an initial set of formant synthesis control parameters can be quite time consuming. Although resynthesis using a formant synthesizer can be manageable for small speech databases or for simple utterances consisting of a syllable or two, the derivation of an initial set of control parameters can be prohibitively time consuming to carry out on a large scale with even moderately complex connected speech databases such as the two sentence databases that were tested in the present study.

There are, on the other hand, some research applications for which the damped sinewave method is poorly

suited. In contrast to formant synthesis, and in common with many other channel vocoders, the DSS uses a source signal—a sequence of spectrally white pulses varying only in amplitude—that differs markedly from the source signal that drives the vocal tract filter in natural speech. Combining the source spectral shape and vocal tract filter shape into a single filter that is derived from the spectral envelope simplifies the design of the vocoder; however, this design makes the synthesizer unsuitable for studying the perceptual effects of all but the simplest source parameters (e.g., Childers & Lee, 1991; Klatt & Klatt, 1990; Rosenberg, 1971).

This is not to suggest that vocal quality is poorly preserved by the DSS. On the contrary, it is our impression that individual speaker identity in general and vocal quality in particular are typically well preserved by the vocoder. The problem is simply that, unlike a formant synthesizer, explicit control of source features is limited to  $F_0$ , periodicity, and source amplitude.

The results of intelligibility tests with the DSS were mixed. Sentence intelligibility was excellent for both the HINT and TIMIT utterances. Vowel identity was quite well preserved by the synthesizer, with overall vowel intelligibility for a large, multispeaker database that was a scant 2.7% lower than that for the original digitized signals and slightly (but reliably) higher than that for formant synthesized versions of the same syllables. Given the nature of the signal processing that is involved in the DSS, we believe that these findings are consistent with the idea that conveying vowel quality does not require a detailed match in spectral shape between the original and synthesized signals. Rather, conveying vowel identity requires reconstructing a signal that matches the original primarily in the spectral-peak regions. This conclusion is consistent with the widely cited findings of Klatt (1982), showing that vowel quality is strongly affected by changes in formant frequencies but not by other spectral shape details, such as spectral tilt, formant bandwidths, and spectral notches. The present results with the /hVd/ signals, however, demonstrate that the spectral peaks that are used to reconstruct the signal need not correspond to labeled formants in the traditional sense. In our view, the findings here and in Klatt (1982) are consistent with a spectral shape model, but one that is strongly weighted toward energy in and around spectral peaks. (For a detailed test of this principle using an independently developed masking method that is strikingly similar to the masked spectral peak representation used here, see Liénard & Di Benedetto, 2000).

Consonant identity was not as well preserved by the synthesizer as vowel identity, with overall consonant intelligibility some 18% lower for the synthesized signals than the original recordings. As discussed above,

consonant transmission loss was especially high for stops and affricates, suggesting that the 64-ms analysis window may have been too large to capture the rapid spectral changes that characterize these sounds. Shorter analysis windows, however, might well be expected to result in less reliable estimates of perceptually relevant aspects of the spectral envelope for vowels and other sounds with spectra that change more slowly over time. We have considered but not yet implemented a rather straightforward modification to the analysis algorithm involving a spectral-change-based adaptive algorithm. This modification sets the analysis window to a relatively small value during portions of the signal undergoing rapid spectral change and, conversely, uses a relatively long window when the spectrum is changing more slowly. As we suggested above, the less-than-perfect transmission of consonant identity might also indicate that the transmission of consonant features requires the preservation of some fine spectral shape details that are not always preserved accurately when the spectral envelope is estimated entirely from a handful of smoothed envelope peaks. We are currently testing this idea with a study that compares the intelligibility of speech produced by the DSS with an otherwise identical synthesizer in which the impulse response is computed directly from the detailed spectrum envelope.

## Acknowledgments

This work was supported by a grant from the National Institutes of Health (2-R01-DC01661) to Western Michigan University. We are grateful to Robert Shannon of House Ear Institute for making the consonant recordings available to us and to Michael Dorman of Arizona State University for providing the HINT sentences. We would also like to thank Michael Clark for helpful comments on an earlier draft.

## References

- Bladon, A.** (1982). Arguments against formants in the auditory representation of speech. In R. Carlson & B. Granstrom (Eds.), *The representation of speech in the peripheral auditory system* (pp. 95–102). Amsterdam: Elsevier Biomedical Press.
- Bladon, A., & Lindblom, B.** (1981). Modeling the judgment of vowel quality differences. *Journal of the Acoustical Society of America*, *69*, 1414–1422.
- Childers, D. G., & Lee, C. K.** (1991). Vocal quality factors: Analysis, synthesis, and perception. *Journal of the Acoustical Society of America*, *90*, 2394–2410.
- Cooper, F. S., Delattre, P. C., Liberman, A. M., Borst, J. M., & Gerstman, L. J.** (1952). Some experiments on the perception of synthetic speech sounds. *Journal of the Acoustical Society of America*, *24*, 597–606.
- Dudley, H.** (1939). Remaking speech. *Journal of the Acoustical Society of America*, *11*, 169–177.
- Hedlin, P.** (1982). A representation of speech with partials. In R. Carlson & B. Granstrom (Eds.), *The representation of speech in the peripheral auditory system* (pp. 247–250). Amsterdam: Elsevier Biomedical Press.
- Hertrich, I., & Ackermann, H.** (1999). A vowel synthesizer based on formant sinusoids modulated by fundamental frequency. *Journal of the Acoustical Society of America*, *106*, 2988–2990.
- Hess, W. J.** (1983). *Pitch determination of speech signals: Algorithms and devices*. Berlin, Germany: Springer-Verlag.
- Hillenbrand, J. M., Getty, L. A., Clark, M. J., & Wheeler, K.** (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America*, *97*, 3099–3111.
- Hillenbrand, J. M., & Nearey, T. N.** (1999). Identification of resynthesized /hVd/ syllables: Effects of formant contour. *Journal of the Acoustical Society of America*, *105*, 3509–3523.
- Indefrey, H., Hess, W. J., & Seeser, G.** (1985). Design and evaluation of double-transform pitch determination algorithms with nonlinear distortion in the frequency domain. *Proceedings of the 1985 International Conference on Acoustics, Speech and Signal Processing*, *2*, 11–12.
- Klatt, D. H.** (1980). Software for a cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America*, *67*, 971–995.
- Klatt, D. H.** (1982). Prediction of perceived phonetic distance from critical-band spectra: A first step. *Proceedings of the 1982 International Conference on Acoustics, Speech, and Signal Processing*, 1278–1281.
- Klatt, D. H.** (1987). Review of text-to-speech conversion for English. *Journal of the Acoustical Society of America*, *82*, 737–793.
- Klatt, D. H., & Klatt L. C.** (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America*, *87*, 820–857.
- Liénard, J.-S., & Di Benedetto, M.-G.** (2000). Extracting vowel characteristics from smoothed spectra. *Journal of the Acoustical Society of America*, *108*(Suppl. 1), 2602.
- McAuley, R. J., & Quatieri, T. F.** (1986). Speech analysis/synthesis based on sinusoidal representation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *22*, 330–338.
- Moulines, E., & Charpentier, F.** (1990). Pitch synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, *9*, 453–467.
- Nilsson, M., Soli, S., & Sullivan, J.** (1994). Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise. *Journal of the Acoustical Society of America*, *95*, 1085–1099.
- Oppenheim, A. V.** (1969). Speech analysis-synthesis system based on homomorphic filtering. *Journal of the Acoustical Society of America*, *45*, 458–465.
- O’Shaughnessy, D.** (2000). *Speech communications: Human and machine*. Piscataway, NJ: IEEE Press.
- Paul, D. B.** (1981). The spectral envelope estimation vocoder. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *29*, 786–794.
- Rodet, X.** (1984). Time-domain formant wave-function synthesis. *Computer Music Journal*, *8*, 9–14.

**Rosenberg, A.** (1971). Effect of glottal pulse shape on the quality of natural vowels. *Journal of the Acoustical Society of America*, *49*, 583–590.

**Shannon, R. V., Jansvold, A., Padilla, M., Robert, M. E., & Wang, X.** (1999). Consonant recordings for phoneme recognition testing. *Journal of the Acoustical Society of America*, *106*, L71–L74.

**Witten, I. H.** (1982). *Principles of computer speech*. London: Academic Press.

**Zahorian, S., & Jagharghi, A.** (1993). Spectral shape features versus formants as acoustic correlates for

vowels. *Journal of the Acoustical Society of America*, *94*, 1966–1982.

---

Received November 5, 2001

Accepted March 22, 2002

DOI: 10.1044/1092-4388(2002/XXX)

Contact author: James M. Hillenbrand, PhD, Department of Speech Pathology and Audiology, Western Michigan University, Kalamazoo, MI 49008.  
E-mail: james.hillenbrand@wmich.edu