Speech perception based on spectral peaks versus spectral shape

James M. Hillenbrand[a)]
Department of Speech Pathology and Audiology, Western Michigan University, Kalamazoo, Michigan 49008

Robert A. Houde
Center for Communications Research, 125 Tech Park Drive, Rochester, New York 14623

Robert T. Gayvert
Gayvert Consulting, 16 Chase View Road, Fairport, New York 14450

This study was designed to measure the relative contributions to speech intelligibility of spectral envelope peaks (including, but not limited to formants) versus the detailed shape of the spectral envelope. The problem was addressed by asking listeners to identify sentences and nonsense syllables that were generated by two structurally identical source-filter synthesizers, one of which constructs the filter function based on the detailed spectral envelope shape while the other constructs the filter function using a purposely coarse estimate that is based entirely on the distribution of peaks in the envelope. Viewed in the broadest terms the results showed that nearly as much speech information is conveyed by the peaks-only method as by the detail-preserving method. Just as clearly, however, every test showed some measurable advantage for spectral detail, although the differences were not large in absolute terms. © *2006 Acoustical Society of America.* [DOI: 10.1121/1.2188369]

I. INTRODUCTION

The most casual inspection of the spectrogram of a naturally spoken utterance reveals a wealth of detail in the evolution of the spectrum over time. It has been clear from the start that some spectral details are more intimately associated with the transmission of phonetic information than others. A good deal of research in phonetic perception has revolved around attempts to draw inferences about the nature of the auditory and perceptual mechanisms that mediate phonetic recognition by conducting listening experiments using carefully controlled speech signals which are contrived in such a way as to retain only some characteristics of the speech signal while purposely removing or distorting other spectral details. The present study was designed to address one aspect of this problem having to do with the relative contributions to speech intelligibility of spectral envelope peaks versus the detailed shape of the spectral envelope. The problem was addressed by asking listeners to identify sentences and nonsense syllables that were generated by two structurally identical source-filter synthesizers, one of which constructs the filter function based on the detailed spectral envelope shape while the other constructs the filter function based entirely on peaks in the envelope.

A closely related question was addressed using very different methods in a series of interconnected experiments on the perception of vowel quality by Carlson, Granstrom, and Klatt (1979), Carlson and Granstrom (1979), and Klatt (1982). Carlson *et al.* (1979) used a harmonic synthesizer to generate 66 signals that were variations on a reference vowel with an /æ/-like quality. The test signals differed from the reference vowel in one of two kinds of parameters: (1) formant frequencies or (2) parameters that affected the detailed shape of the spectrum *but did not involve manipulation of the formant frequencies of the reference vowel* (e.g., high- or low-pass filtering, spectral tilt, overall amplitude, formant bandwidths, spectral notches). Subjects were asked to judge the *overall psychophysical distance* between each test signal and the reference vowel. Results showed that substantial differences in sound quality could be induced either by altering the formant frequency pattern or by altering many of the nonformant-related spectral shape features, especially those affecting the shape of the spectrum in the low frequencies. Carlson and Granstrom (1979) went on to show that these psychophysical distance judgments could be predicted based on processing schemes that simulated the characteristics of low-level auditory analysis (see also Lindblom, 1978; Bladon and Lindblom, 1981). In a widely cited follow-up study, Klatt (1982) argued that applying findings such as these to phonetic perception required a listening task in which subjects are asked to rate *phonetic* differences among stimuli, ignoring other timbre differences as much as possible. Klatt presented the same 66 /æ/-like signals from Carlson *et al.* (1979) to a new group of listeners, but the results from the phonetic-distance task were dramatically different from the earlier findings using the psychophysical distance task. Klatt reported that, "… only formant frequency changes induced large changes in phonetic distance. Even though filtering and spectral tilt conditions produce substantial changes in the spectrum, these changes are apparently ignored when making phonetic judgments" (p. 1278). Similar conclusions were

reached in tests using both voiced and whispered vowels with qualities similar to /ɑ/. Significantly, differences in the potency of many spectral shape cues in signaling phonetic versus overall psychophysical changes were often quite large. For example, the effect of a 400 Hz high-pass filter on psychophysical distances was nearly 30 times greater than the effect of that same manipulation on phonetic distances. Although these results might be seen as favoring a formant tracking mechanism, Klatt argued that formant tracking was implausible on a number of grounds (see also Bladon, 1982; Zahorian and Jagharghi, 1993). Klatt contended that his findings were best explained by assuming a spectral-shape pattern matching scheme which is very sensitive to spectral peaks but relatively insensitive to other aspects of the spectrum. A distance metric was developed based on weighted differences in spectral slope, with greater weights being assigned to spectral differences at or near spectral peaks. The weighted slope measure provided accurate predictions of the listener-derived phonetic distance measures.

The present experiment was designed to extend the insights of Klatt's (1982) experiment on the relative contributions of formants versus spectral-shape details. Klatt's evidence that formants are far more important than other aspects of the spectrum is clear enough. However, Klatt's findings are limited to vowel quality, and even at that they are based on results from just two stylized, static vowels. In the present experiment we used methods quite different from Klatt's to address the same sort of question, but we extended the test material to sentences, vowels in /hVd/ syllables, and consonants in CV syllables. The method involved measuring the intelligibility of sentences and nonsense syllables produced by two source-filter synthesizers. The two synthesizers differed only with respect to the methods that were used to estimate the filter function. For a spectral envelope synthesizer (SES), the filter was estimated by making a fine-grained measurement of the spectral envelope of the speech signal. For a damped sine wave synthesizer (DSS), on the other hand, a purposely coarse estimate was used that was based entirely on the distribution of peaks in the spectral envelope. Klatt's findings on vowel color would lead one to expect that vowel identity would be conveyed nearly as well by the coarse, peaks-only damped sine-wave synthesizer as by the fine-detail-preserving spectral-envelope synthesizer. However, Klatt's findings provide no basis for predicting how well consonant identity and sentence intelligibility will be preserved by these two synthesis methods.

II. METHOD

A. Speech Material

1. Vowel database

Vowels in /hVd/ context were selected from recordings made by Hillenbrand, Getty, Clark, and Wheeler (1995). The full 1668-syllable database consists of 12 vowels (/i,ɪ,e, ɛ,æ,ɑ,ɔ,o,ʊ,u,ʌ,ɝ/) spoken by 45 men, 48 women, and 46 10- to 12-year-old children. A 300-utterance subset of this database was selected from the larger database, consisting of 25 tokens of each vowel of the 12 vowels, with at least one token from 123 of the 139 talkers, and roughly equal num-

bers of tokens spoken by men, women, and children (see Hillenbrand and Nearey, 1999, for additional details). The /hVd/ syllables were unmodified from their original 16 kHz sample rate.

2. Consonant database

Consonant intelligibility was tested using signals drawn from the Shannon, Jensvold, Padilla, Robert, and Wang (1999) database. The full database consists of CV, VC, and symmetrical VCV syllables formed by 25 consonants and 3 vowels (/ɑ,i,u/) spoken by five men and five women. A 276-syllable subset of the full database was chosen for the present study, consisting of CV syllables only formed by 23 consonants (/b,d,g,p,t,k,m,n,l,r,f,v,θ,ð,s,z,ʃ,ʒ,tʃ,dʒ,j,w,h/) with all three vowels spoken by two men and two women. The test signals, which were originally recorded at 44.1 kHz, were digitally lowpass filtered at 7.2 kHz and down sampled to 16 kHz. Pilot testing showed that 37 of the 276 syllables were not well identified by listeners, so these signals were omitted, leaving 239 syllables.

3. Sentences

Subjects were also tested on a sentence transcription task using utterances drawn from two databases. One database consisted of the 250 sentences used in the Hearing In Noise Test (HINT) described by Nilsson, Soli, and Sullivan (1994; see also Bench, Kowal, and Bamford, 1979). These utterances are syntactically and semantically simple (e.g., "Big dogs can be dangerous.") and are carefully spoken by a single male talker. A second sentence test, which was expected to be more difficult, consisted of 50 utterances drawn from the TIMIT continuous speech database (Garafolo et al., 1993). 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 HINT sentences were digitally low-pass filtered at 7.2 kHz and down sampled to 16 kHz from the original 20.161 kHz sample rate. The TIMIT sentences were unmodified from their original 16 kHz sample rate.

B. Spectral-envelope synthesizer (SES)

1. Design principles

This source-filter synthesizer has some features in common with the spectral envelope estimation vocoder described by Paul (1981). Figure 1 shows the source signal and frequency response curve that would be used to generate a monotone, sustained, phonated /ɑ/. The source signal consists of a sequence of single-sample pulses whose period is determined by the instantaneous fundamental period of the signal that is being reconstructed. A whispered vowel (or any other unvoiced segment) can be synthesized by replacing the periodic pulse train that is shown in Fig. 1 with a sequence of single-sample pulses whose amplitudes are either zero or nonzero, with a probability of 0.5 at each sample point. (A pulse sequence such as this is spectrally indistinguishable from the Gaussian white noise sequence that is commonly used in source-filter synthesizers such as Klatt, 1980.) Mixed-source signals consisting of both periodic and aperi-
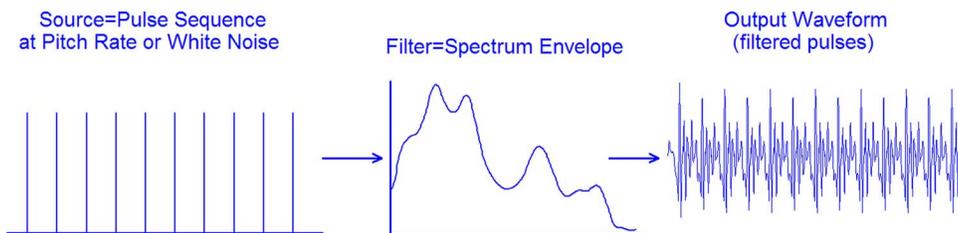
FIG. 1. (Color online) Illustration of synthesis of a sustained vowel using the spectral envelope synthesizer. The source is a sequence of single-sample, spectrally white pulses spaced at the fundamental period. The filter is the spectrum envelope measured from the speech signal being synthesized. A whispered vowel can be synthesized simply by replacing the periodic source signal with a sequence of pulses spaced at random intervals.

odic components (e.g., voiced fricatives or breathy vowels) can be generated simply by adding periodic and random pulse sequences with any desired voiced/unvoiced mixing ratio. The amplitude contour of the source signal is controlled by amplitude modulating the mixed periodic/aperiodic source by the measured amplitude contour of the signal that is being reconstructed. Any number of methods might be used to derive the fundamental frequency (F_0), voiced/unvoiced mixing ratio, and amplitude contours that are needed to construct the source signal through an analysis of a naturally spoken utterance. The source analysis methods that we adopted will be described below.

The filter function that is shown in Fig. 1 is simply the spectrum envelope of the speech signal, which in turn is used to design a finite impulse response (FIR) digital filter to modify the spectral shape of the flat-spectrum source signal. As with the measurement of F_0 and voiced/unvoiced ratio, there are many methods that might be employed to derive the spectrum envelope, including the commonly used methods based on linear predictive coding and the cepstrum. The method used here, which is described below, is a variation on a technique used by Paul (1981).

Figure 1 illustrates a very simple implementation of a spectral envelope synthesizer that would be used to generate a sustained, monotone vowel. Extending the static synthesis illustrated in this figure to a dynamic method that would be needed to reproduce a naturally spoken utterance simply involves updating the relevant source (F_0, voiced/unvoiced ratio, and overall amplitude), and filter (the spectrum envelope) control parameters at some reasonable interval—every 10 ms in our implementation.

2. Source signal (common to SES and DSS)

Generation of the source signal requires the estimation, for each 10 ms speech frame, of three parameters: (1) instantaneous F_0; (2) degree of periodicity (used to set the voiced/unvoiced mixing ratio); and (3) overall amplitude. The methods used to derive these parameters are described in somewhat greater detail in Hillenbrand and Houde (2002). A summary will be presented here. Note that the design principles described in Hillenbrand and Houde are unchanged, although some of the parameter settings were modified, based largely on the listening test results reported in that study.[1] Also, note that exactly the same methods were used to derive the source signal for both the SES and DSS.

F_0 and degree of periodicity are computed from a double-transform method conceptually similar to the cepstrum. The key signal processing steps are illustrated in Fig.

2. The first transform is a 1024-point (64 ms) Hamming-windowed Fourier spectrum, using linear amplitudes. The spectrum is then lightly compressed by raising the spectral amplitudes to the 0.7th power. (The setting for this and other parameters in the F_0 and periodicity analysis were determined through extensive trial-and-error testing. With the notable exception of window sizes for spectrum analyses—discussed below—our informal impression is that the method is not terribly sensitive to modest changes in parameter settings.) The next step is to compute a threshold function as the 110 Hz Gaussian-weighted running average of spectral amplitudes.[2] The threshold function is subtracted from the
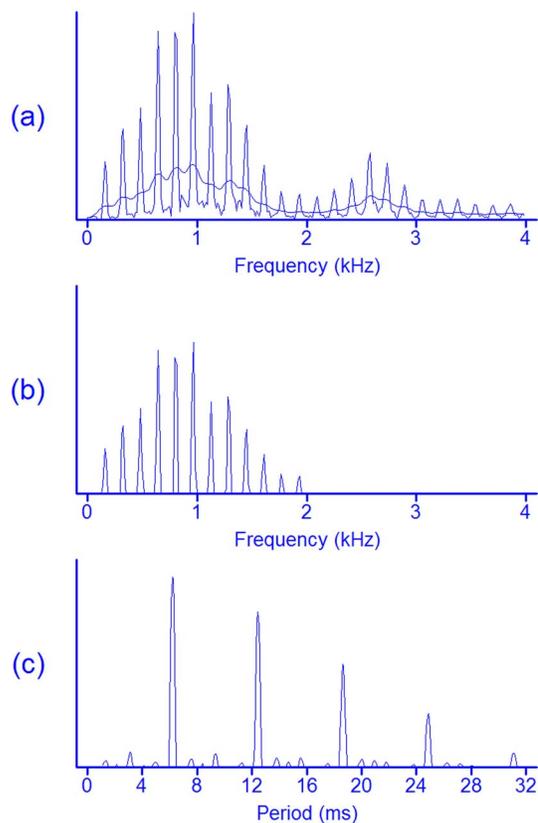
FIG. 2. (Color online) The key signal processing steps used to compute F_0 and degree of periodicity. Panel *a* shows a narrow band FFT after light compression (amplitudes are raised to the 0.7th power). The smooth curve is a threshold function computed as the Gaussian-weighted running average of spectral amplitudes computed over a 110 Hz window. Panel *b* shows the spectrum after subtraction of the threshold function, with negative values set to zero. This step enhances harmonics at the expense of interharmonic noise. Channels below 50 Hz or above 2000 Hz are zeroed out. Panel *c* shows the cosine transform (positive values only) from which fundamental period and degree of periodicity are derived.

compressed spectrum, with spectral values below the threshold set to zero. The purpose of the thresholding operation is to emphasize voice-source harmonics at the expense of interharmonic spectral noise (see Hillenbrand and Houde, 2003).[3] Values in this spectrum below 50 Hz or above 2000 Hz are zeroed out and a cosine transform is computed. The fundamental frequency is defined as the inverse of the shortest period in the cosine transform (estimated with parabolic interpolation) that is at least 70% of the amplitude of the absolute maximum, with the search constrained to the range between 1.25 and 31 ms (32–800 Hz). F_0 is measured every 10 ms speech frame, regardless of the degree of signal periodicity.

A degree-of-periodicity ratio is computed for each frame as the amplitude of the largest peak in the cosine transform divided by the sum of the amplitudes in the spectrum from which the cosine transform was computed. This ratio is equivalent to the sum of the amplitudes of all harmonics of the fundamental (including the fundamental) divided by the sum of the amplitudes of all components in the spectrum from which the cosine transform is computed. The voiced/unvoiced mixing ratio is derived from the periodicity ratio by a nonlinear mapping in which: (a) values below 0.15 are set to 0.0; (b) values above 0.75 are set to 1.0; and (c) values between 0.15 and 0.75 are linearly scaled between 0.0 and 1.0.

The final measurement that is needed to generate the source signal is the amplitude contour of the utterance being synthesized, which is computed as the Gaussian-weighted running average of the full-wave rectified time wave form, computed over a 20 ms window. 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 described above. 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, while 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. Prior to any other scaling, the ratio of the peak amplitude of the aperiodic pulses to that of the periodic pulses is set to 0.33, resulting in similar subjective loudnesses for the relatively sparse periodic sequences and the more dense aperiodic pulse sequences. The periodic pulse sequence is then amplitude modulated by the voiced/unvoiced mixing function while the aperiodic pulse sequence is amplitude modulated by its complement. The periodic and aperiodic wave forms are then mixed and the sum is amplitude modulated by the amplitude contour measured from the original signal.

3. Filter function

In terms of the experimental goals of the study, the key feature of the spectral envelope synthesizer is that the filter function is constructed from a fine-grained analysis of the spectral envelope. The filter function is derived from the spectral envelope using a method which we have called the *harmonic envelope*, inspired by a related technique devel-
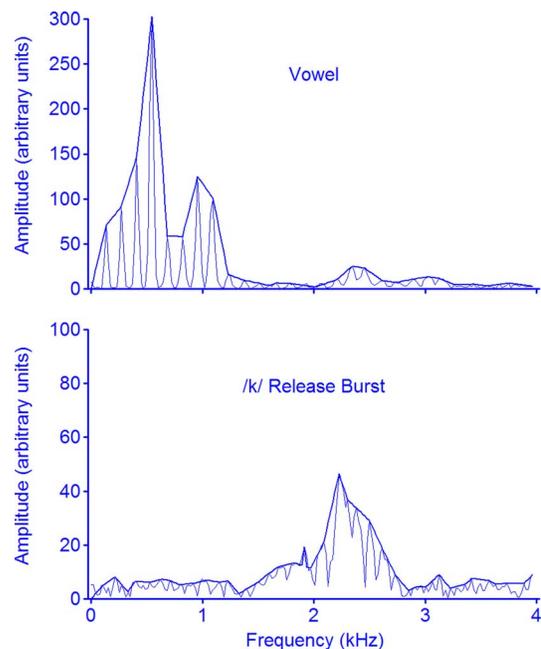
FIG. 3. (Color online) The harmonic envelope for a vowel (top panel) and a consonant release burst (bottom panel).

oped by Paul (1981). For each analysis frame, the method begins with the calculation of a narrow band Fourier spectrum, using a 512-point (32 ms) Hamming window, and with an estimate of F_0 (using the method described above). The F_0 estimate is used to locate individual harmonic peaks in the Fourier spectrum. The first harmonic is defined as the highest amplitude peak in the range beginning at $F_0 \cdot 0.5$ and extending to $F_0 \cdot 1.5$; i.e., the search window is centered at the fundamental frequency and extends upward and downward by one half of the fundamental. Similarly, the second harmonic is defined as the highest amplitude peak in a window centered at $F_0 \cdot 2.0$, plus or minus one half the fundamental. The search for harmonic peaks continues until the Nyquist frequency is reached. Once all harmonic peaks have been located, the remainder of the spectrum envelope is computed simply by linearly interpolating between harmonics. The result is an envelope with the same frequency resolution as the original spectrum (256 points, or 31.25 Hz per channel). The top panel of Fig. 3 shows an example of a harmonic envelope for a voiced speech segment (the spectrum is shown up to 4 kHz only). Despite our use of the term *harmonic envelope*, the method makes no distinction between periodic and aperiodic speech segments. Recall that F_0 is measured for all speech frames, regardless of the degree of periodicity. These F_0 estimates are used to define the envelope for unvoiced and marginally periodic segments using the method just described, in spite of the fact that the peaks will often not correspond to harmonics. The lower panel of Fig. 3 shows an example of a harmonic envelope for a spectral slice taken from a stop release burst for a /k/.

The 256 amplitudes in the harmonic envelope are used directly to define the gain function of a finite-impulse response filter, which is computed as the inverse Fourier transform of the harmonic envelope, resampled to 256 points (16 ms). The phase response of the filter is set such that
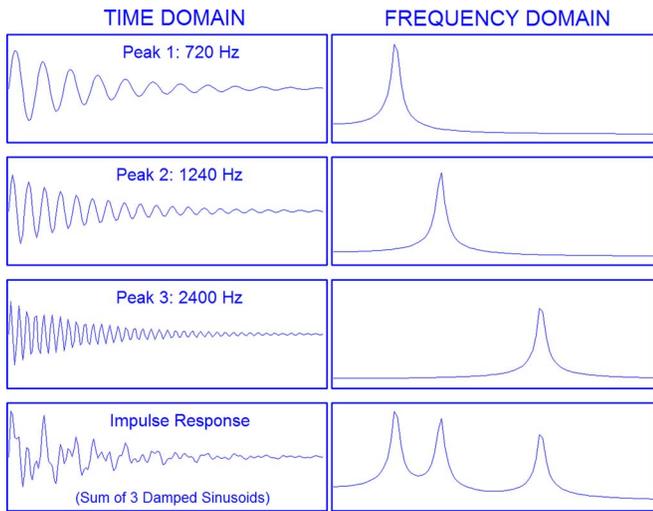
FIG. 4. (Color online) Synthesis of a sustained vowel by summing exponentially damped sine waves.

successive components alternate between 0° and 180° (i.e., the first component is set to 0°, the second component is set to 180°, the third component is set to 0°, and so on). This phase pattern results in an impulse response that reaches peak amplitude at the middle of the window and approaches zero at the endpoints (see Hillenbrand and Houde, 1996, for a discussion). The final step in the calculation of the filter function is to scale the peak amplitude in the impulse response for each frame to a constant. As with all other control parameters, the envelope-derived impulse response is updated every 10 ms. Once the source signal and frame-by-frame sequence of impulse responses have been computed, synthesis is simply a matter of convolving the source signal with the time-varying impulse response.

C. Damped sine wave synthesizer (DSS)

A full description of the DSS, based on an implementation similar but not identical to the one used here, can be found in Hillenbrand and Houde (2002). The source signal for the DSS is identical to the one that is generated for the SES; that is, it consists of a sequence of single-sample pulses spaced at the fundamental period for voiced speech, at random intervals for unvoiced speech, or mixed periodic and random intervals for mixed-source signals. Also in common with the SES, the speech signal is synthesized by convolving this source signal with the time-varying impulse response of a digital FIR filter. The only difference between the two synthesizers is that the DSS uses a purposely coarse method for specifying the filter function in which the impulse response is the sum of exponentially damped sine waves at frequencies and amplitudes corresponding to peaks that are extracted from the spectrum envelope. The logic underlying the method is illustrated in Fig. 4, which shows the generation of an impulse response for an /ɑ/-like vowel with envelope peaks at 720, 1240, and 2400 Hz. The individual components that are summed to create the impulse response are damped sinusoids of the form

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

where a is the amplitude, t is the time (s), f is the frequency (Hz), b is the bandwidth (Hz). As shown in the frequency-domain representations to the right of Fig. 4, each damped sinusoid has a well-defined formant-like peak at the frequency of the sinusoid. When the individual damped sinusoids are summed, the result defines the impulse response of a filter having resonances corresponding to the sinusoidal frequencies.

The frequencies and amplitudes of the damped sine waves are measured from smooth spectra of the signal being synthesized. In the implementation described here, the bandwidths are fixed at 80 Hz. The spectral peaks can be measured from any type of smooth spectrum. The method used in the present study begins with the same harmonic envelope that is used in the SES. The harmonic envelope is then smoothed[4] by a Gaussian-weighted running average with a window size equal to the instantaneous fundamental frequency, measured using the technique described above (Fig. 5, panels a and b). A thresholding procedure is then used to suppress minor spectral peaks. A threshold function is computed as the 800 Hz Gaussian-weighted running average of the smoothed spectrum. The threshold function is subtracted from the smoothed envelope and all spectral amplitudes below the threshold are set to zero [Fig. 5, panels (c) and (d)]. Peak amplitudes are extracted from the smoothed envelope prior to thresholding [panel (c)] and peak frequencies are extracted from the spectrum derived from the thresholding operation [panel (d)]. No continuity constraints are used and no limit is placed on the number of spectral peaks per frame, which averaged just under 10 for the 50 TIMIT sentences (bandwidth=8 kHz). The signal processing steps are shown in spectrographic form in Fig. 6 for the TIMIT sentence, "If dark came they would lose her." Of special note in this figure is the bottom display showing the sequence of envelope peaks that is used to derive the filter function for the DSS. It can be seen that the formant structure is reasonably well preserved in some segments of the utterance but quite poorly preserved in others. For example, note the large number of peaks in the word "came" that do not correspond to formants. As will be noted below, this utterance, along with other sentences in the TIMIT and HINT databases, was highly intelligible.

For each spectral peak that is detected using these methods, a damped sinusoid is generated at the measured frequency and amplitude and with a fixed bandwidth of 80 Hz. The damped sinusoids for each frame are then summed and the sum of all damped sinusoids for the frame is scaled to a constant peak amplitude. This sum serves as the finite impulse response which defines the filter component of the source-filter synthesizer. A fixed FIR length of 256 points (16 ms) is used for all frames. The final synthesis step is simply the convolution of the sequence of time-varying FIRs with the source signal described above. Additional technical details about the DSS can be found in Hillenbrand and Houde (2002).

J. Acoust. Soc. Am., Vol. 119, No. 6, June 2006

Hillenbrand *et al.*: Spectral peaks versus spectral shape   4045
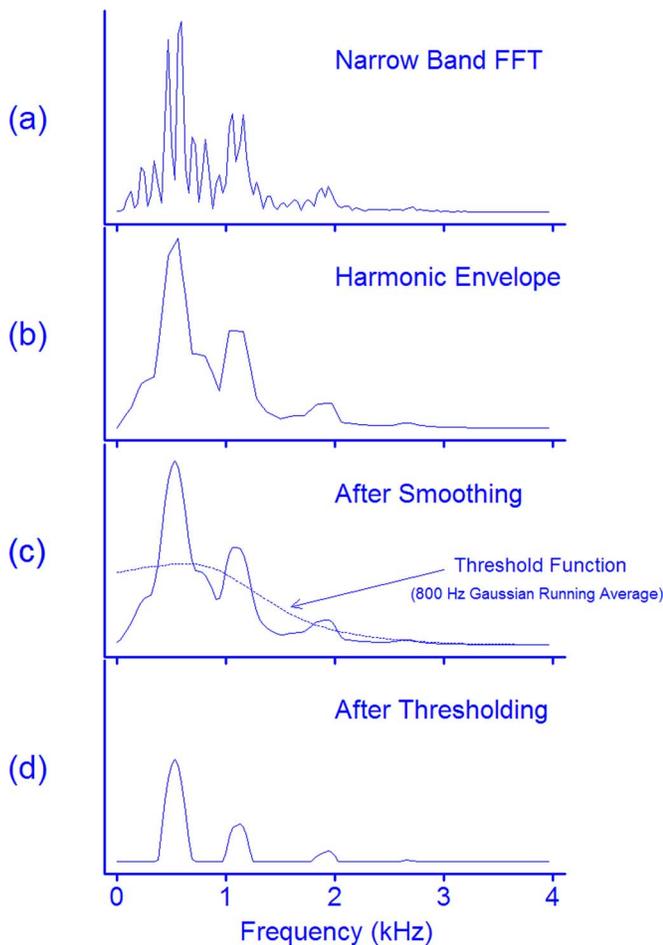
FIG. 5. (Color online) Spectral slice showing the signal-processing steps used in the extraction of spectral peak frequencies and amplitudes for the damped sine wave synthesizer: (a) narrow band Fourier spectrum, computed over a 32-ms Hamming window; (b) harmonic envelope computed by linear interpolation between harmonic peaks; (c) smoothed harmonic envelope (peak amplitudes are measured from this spectrum) with threshold function computed as the 800 Hz Gaussian-weighted running average of spectral amplitudes; and (d) spectrum after subtraction of the threshold function, with spectral values below the threshold set to zero (peak frequencies are extracted from this spectrum). Spectra are shown up to 4 kHz only.

Summary

The SES and DSS are structurally identical synthesizers that use the same source signal and the same source-filter convolution algorithm. The SES and DSS differ only in the method that is used to estimate the filter: The SES attempts to preserve the envelope shape as faithfully as possible by defining the filter as the envelope of the input signal, measured at 256 discrete frequencies (i.e., with a precision of 31.25 Hz). The DSS, on the other hand, uses a purposely coarse filter whose shape is determined by the frequencies and amplitudes of broad envelope peaks, which average roughly one peak per 1000 Hz. Below it will be demonstrated that the fine-detail-preserving SES produces considerably more natural sounding speech than the DSS. The central experimental question concerns the extent to which this preservation of fine spectral detail is also associated with an increase in speech intelligibility.
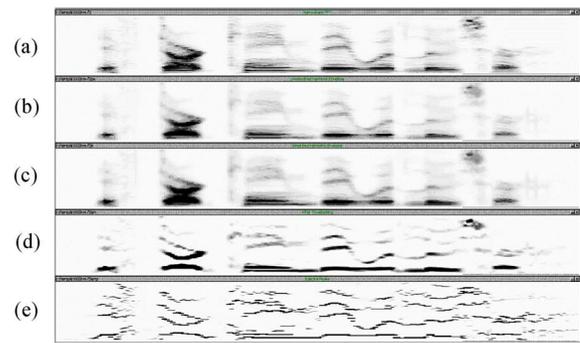
FIG. 6. Spectrograms showing the signal-processing steps used in the extraction of spectral peak frequencies and amplitudes for the damped sine wave synthesizer. The utterance is "If dark came they would lose her," from the TIMIT database. From top to bottom: (a) narrow band spectrogram; (b) harmonic envelope; (c) smoothed harmonic envelope; (d) envelope after subtraction of 800 Hz Gaussian-weighted running average, with negative values set to zero; and (e) sequence of peaks extracted from spectra after subtraction of the running average. Spectrograms are shown up to 4 kHz only.

D. Listeners

Listeners were undergraduate and graduate students in the Speech Pathology and Audiology program at Western Michigan University. The listeners had normal hearing sensitivity, had completed an introductory course in phonetics, and were either paid or given course credit for their participation. Separate groups of listeners participated in one of four experiments: (1) sentence intelligibility using the HINT sentences ($N=18$); (2) sentence intelligibility using the TIMIT sentences ($N=19$); (3) vowel intelligibility using /hVd/ syllables drawn from Hillenbrand *et al.* (1995) ($N=13$); and (4) consonant intelligibility using CV syllables from Shannon *et al.* (1999) ($N=25$). Twelve additional listeners were recruited for a preliminary experiment involving judgments of speech naturalness (described below).

E. Procedures

1. General

Stimulus presentation, displays and, where appropriate, the collection of subject responses were controlled by a general-purpose experiment design and control program (Hillenbrand and Gayvert, 2005). The signals to be used in each experiment were scaled to a common rms intensity (the highest rms value that could be achieved without any of the signals exceeding the 16-bit limit of the digital-to-analog converter). The signals were low-pass filtered at 7.2 kHz, amplified, and presented free field at peak intensities averaging 75 dBA using a Paradigm Titan loudspeaker positioned about 1 m from the listener.

2. Sentence intelligibility

The 250 sentences from the HINT database were presented in random order to listeners, who were asked to repeat the sentence to an experimenter seated adjacent to the subject. On a random half of the trials the SES version of the utterance was presented and on the other half the DSS version was presented. Listeners were given the option of replaying the sentence once or twice before responding. The

experimenter scored the subject's response, keeping a record of the number of times the listener requested a repetition of the sentence, then pressed a key to advance to the next trial. The same procedure was used for the 50 TIMIT sentences.

3. Vowel intelligibility

Listeners identified four versions of each of the 300 /hVd/ utterances from the Hillenbrand *et al.* (1995) database: (1) the naturally spoken utterance; (2) a SES synthesized version; (3) a DSS synthesized version; and (4) a formant-synthesized version generated with the Klatt and Klatt (1990) synthesizer. Formant-synthesized signals were used for the vowel intelligibility tests only since the test signals were available from a previous study (Hillenbrand and Nearey, 1999). Formant-synthesis control parameters were derived from hand-edited F_0 and formant contours measured in Hillenbrand *et al.* (1995). Five formants were used, the synthesizer was run in cascade mode (meaning that formant amplitudes were not explicitly controlled), and formant bandwidths during the vowel portion of the /hVd/ utterances remained at their default values $(B_1=90, B_2=110, B_3=170, B_4=400, B_5=500)$. A detailed description of the formant-synthesized test signals can be found in Hillenbrand and Nearey. The 1200 syllables were presented in a single random order, scrambled separately for each listener. Other aspects of the testing procedure were analogous to those used for the consonant intelligibility tests.

4. Consonant intelligibility

Listeners identified three versions of each of the 239 CV syllables from Shannon *et al.* (1999): (1) the naturally spoken utterance; (2) a SES synthesized version, and (3) a DSS synthesized version. The 717 syllables were presented in a single random order and scrambled separately for each listener. Listeners identified the initial consonant by clicking 1 of 23 buttons labeled with both phonetic symbols and key words. Listeners were given the option of replaying an utterance as many times as desired before responding.

5. Speech naturalness

Although the primary focus of this study was speech intelligibility, a preliminary experiment was conducted involving judgments of speech naturalness. Our informal impression was that speech synthesized with the SES method was more natural sounding than that produced with the DSS. This would be consistent with the underlying premise that the SES method preserves more of the fine spectral detail of the original signal than the DSS method. Consequently, an experiment was run to formalize our subjective impressions about these differences in naturalness. Using the 50 TIMIT sentences, listeners heard two utterances on each trial. The naturally spoken sentence was played first, followed by either the SES or DSS version of the same sentence, determined randomly, with a probability of 0.5, and scrambled separately for each listener. Listeners were told that the first sentence was a recording of an original, naturally spoken utterance, while the second was a version of the same sentence that had been created artificially by a computer speech synthesizer. Listeners were asked to judge the naturalness of the synthesized sentence in relation to the original sentence. They were told that the original sentence had a naturalness value of 100 by definition, so if the synthesized utterance sounded half as natural as the original they were to assign a value of 50, if it sounded 1/10th as natural, they were to assigned it a value of 10, etc.

III. RESULTS

A. Naturalness

Average naturalness ratings were significantly higher for the SES signals (81.8) than for the DSS signals (59.4; $t=3.78$, $df=11$, $p<0.05$). These findings confirm an important assumption underlying the study; namely, that the SES method does a better job of preserving the detailed shape of the time-varying spectrum than does the peaks-only DSS method. It is clear, then, that preserving spectral shape detail plays a significant role in controlling the overall quality of the synthesized utterance. The key question is whether this advantage for the SES in the preservation of spectral shape detail is accompanied by a corresponding advantage in the transmission of *phonetic* information.

B. Sentence intelligibility

Intelligibility for the HINT sentences, measured as the percentage of content words correctly repeated by the listener (following Nilsson *et al.*, 1994), was slightly higher for the SES (99.8%) signals than the DSS signals (98.7%). However, variability across listeners was quite low and each of the 18 individual listeners showed higher intelligibility scores for the SES signals. Consequently, despite the small difference between the means, the difference in intelligibility for the two synthesis conditions (using arcsine transformed percent correct values) was highly significant ($t=7.6$, $df=17$, $p<0.001$). Listeners rarely asked for either version of the HINT sentences to be replayed. However, they were far more likely to request a repetition of the DSS version (median=7 per session) than the SES version (median=1). A Wilcoxon signed-rank test on these non-normally distributed data showed that this difference was highly significant ($z=3.5$, $p<0.01$).

Results for the TIMIT sentences were similar. Intelligibility was slightly but significantly higher for the SES (96.9%) signals than the DSS signals (95.7%; $t=1.7$, $df=18$, $p<0.05$). Further, listeners requested significantly more repetitions of the DSS versions (median=7) than the SES versions (median=5; $z=3.2$, $p<0.01$).

C. Vowel intelligibility

Vowel intelligibility results for the natural signals and for the three types of synthesized signals (SES, DSS, and formant synthesis), averaged across the 13 listeners, are shown in Table I. It can be seen that vowel intelligibility was nearly identical for the natural and SES signals, but there was a drop in intelligibility of some 5–6 % for either the DSS or formant-synthesized signals. The DSS and formant-synthesized signals produced very similar recognition rates.

J. Acoust. Soc. Am., Vol. 119, No. 6, June 2006

Hillenbrand *et al.*: Spectral peaks versus spectral shape   4047

TABLE I. Vowel intelligibility for four different versions of the 300 /hVd/ utterances from Hillenbrand *et al.* (1995): (1) naturally spoken (NAT), (2) synthesized with the spectral envelope synthesizer (SES), (3) synthesized with the damped-sine wave synthesizer (DSS), and (4) synthesized with the Klatt and Klatt (1990) formant synthesizer (Klsyn).

NAT	SES	DSS	Klsyn
Mean (sd)	Mean (sd)	Mean (sd)	Mean (sd)
95.2 (2.9)	95.0 (2.6)	89.3 (5.0)	90.0 (3.1)

A two-way repeated measures analysis of variance (ANOVA) computed on arcsine transformed percent correct scores showed highly significant effects for Stimulus Type (natural/SES/DSS/formant synthesis) $[F (3,51)=56.3,\ p<0.001]$ and Vowel Category $[F (11,187)=6.2,\ p<0.001]$ as well as a significant interaction $[F (33,561)=7.7,\ p<0.001]$. Bonferroni *post hoc* tests showed: (1) no difference between the natural and SES signals; (2) no difference between DSS and formant-synthesized signals; and (3) differences between both the natural and SES signals versus the DSS and formant-synthesized signals. The main effect for Vowel is unsurprising in light of the many studies show that some vowels are more readily identified than others (e.g., Peterson and Barney, 1952; Hillenbrand *et al.*, 1995). As can be seen in Fig. 7, the effect of stimulus type was not uniform across vowels, resulting in the Stimulus-Type-by-Vowel interaction. Analysis and discussion of variations across vowels for natural versus formant-synthesized vowels can be found in Hillenbrand and Nearey (1999). Other aspects of the interaction patterns shown in Fig. 7 are rather complex and no simple explanation seems apparent to us.

The full confusion matrices for each of the four conditions are shown in Tables A1–A4. The main point worth noting in these tables is that the errors in all four matrices are unremarkable, showing the usual confusions that are seen among adjacent vowels. This is hardly surprising for the naturally spoken vowels, or for the SES vowels given that the goal was to preserve the detailed spectral shape of the original signal as closely as possible. Equally unsurprising is the well-behaved confusion matrix for the formant synthesized signals since these were generated from hand-edited formant contours. It is worth noting, however, that the confusion matrix for the DSS signals is also dominated by confusions among adjacent vowels, despite the fact that these signals were generated from unedited spectral peaks rather than hand-edited formants.

The main conclusion to be drawn from the vowel intelligibility tests is that the SES method, which preserves the fine details of the original spectrum, conveys vowel identity almost perfectly while the peaks-only methods (DSS and formant synthesis) result in a modest (5%–6%) but significant loss of information conveying vowel identity. A very similar drop in vowel intelligibility was reported by Hillenbrand and Nearey (1999) in a comparison of naturally spoken (95.4%) and formant synthesized vowels (88.5%). Also significant, and closely related to the point above regarding the general look of the confusion matrices, is the fact that the formant-synthesized signals, which were driven by carefully hand-edited formant tracks, were no more intelligible than the
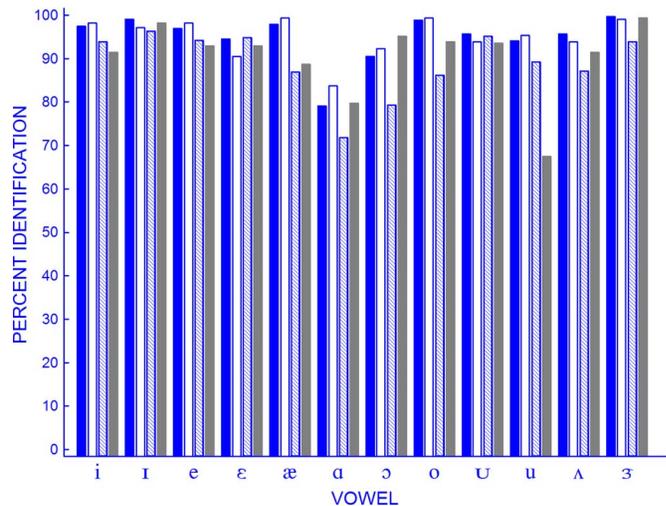
FIG. 7. (Color online) Percent identification as function of vowel category and stimulus type. Going from left to right in each group of bars, the conditions are naturally spoken signals (filled), spectral envelope synthesis (open), damped sine wave synthesis (hatched), and formant synthesis (shaded).

DSS signals, whose spectrum envelopes were created from unedited spectral peaks which sometimes corresponded to formants and sometimes did not. We will return to this point in Sec. IV.

D. Consonant intelligibility

Consonant recognition rates, averaged across all subjects and all consonants, were 97.4% ($sd=1.7$) for the naturally spoken signals, 88.9% ($sd=3.0$) for the SES signals, and 85.9% ($sd=3.3$) for the DSS signals. Although our primary purpose was to compare the fine-detail-preserving SES representation with the peaks-only DSS method, it is clear that the largest effect that was observed was the drop in intelligibility for either of the synthesized signals as compared with the naturally spoken utterances. A two-way repeated measures ANOVA (using arcsine transformed percent correct scores) showed highly significant effects for Stimulus Type $[F (2,48)=414.1,\ p<0.001]$ and Consonant $[F (22,528)=36.2,\ p<0.001]$ as well as a significant interaction $[F (44,1056)=15.9,\ p<0.001]$. Bonferroni *post hoc* tests showed statistically reliable differences among all pair-wise comparisons of Stimulus Types. It is noteworthy, in our view, that the SES-DSS difference, while statistically significant, is a relatively small effect in absolute terms, amounting to an average of just seven additional signals correctly identified (from the total of 239 signals) under the SES condition versus DSS. This effect is also substantially smaller than the ~10 percentage point difference separating the naturally spoken signals from the SES and DSS signals.

The interaction patterns, which are rather complex, are summarized in Table II, which lists the consonants that showed decrements in intelligibility of 10% or greater for three different comparisons. (Full confusion matrices are given in Tables B1–B3). The first group of rows compares identification rates for naturally spoken signals with SES versions of the same signals, the middle group of rows compares naturally spoken signals with DSS versions, and the

TABLE II. Listing of the consonants that showed the largest decrements in intelligibility when comparing: (a) naturally spoken utterances with their spectral envelope synthesis counterparts (the first group of rows); (b) naturally spoken utterances with their damped sinewave synthesis counterparts (the second group of rows); and (c) the spectral envelope signals with their damped sinewave synthesis counterparts (the last group of rows). The last column lists the speech sounds with which the target sound was most commonly confused, with the percentage of such confusions given in parentheses, rounded to the nearest integer.

Speech sound	Natural	SES	Difference	Confused primarily with
b	96.7	41.2	55.5	v(50)
d	100.0	80.7	19.3	ð(15)
k	100.0	81.0	19.0	p(14)
θ	93.1	77.8	15.3	f(14), h(3)
tʃ	97.2	83.1	14.1	ʃ(16)
Speech sound	Natural	DSS	Difference	Confused primarily with
f	91.4	62.1	29.3	θ(36)
ʒ	96.7	73.9	22.8	dʒ (8), ð(7)
b	96.7	74.0	22.7	v(13) ð(5) d(2)
v	98.5	77.8	20.7	ð(10), l(5), m(5)
d	100.0	80.5	19.5	ð(13)
tʃ	97.2	78.5	18.7	ʃ(16), t(3)
r	98.0	80.9	17.1	b(7), w(4), m(4), l(3)
k	100.0	84.5	15.5	p(10)
dʒ	94.2	79.2	15.0	tʃ(7), t(5)
ʃ	98.7	84.4	14.3	s(6), tʃ(6)
g	96.0	83.9	12.1	d(7), k(3)
h	94.8	83.7	11.1	p(11)
Speech sound	SES	DSS	Difference	Confused primarily with
f	91.5	62.1	29.4	θ(36)
v	99.0	77.8	21.2	ð(10), l(5), m(5)
ʒ	88.7	73.9	14.8	ð(7), j (5), z(3)
ʃ	97.3	84.4	12.9	s(6), tʃ(6)
r	91.6	80.9	10.7	b(7), w(4), m(4)

bottom group of rows compares SES and DSS signals. For the natural versus SES comparisons, all of the sounds showing large decrements in intelligibility (/b,d,k,θ,tʃ/) are cued primarily by rapid changes in the spectrum.[5] The single most striking feature of this table is the 55.5% decrement in the intelligibility of /b/, which was heard almost exclusively as /v/ when synthesized by the SES. The most plausible explanation for the strong tendency of SES versions of /b/ to be heard as /v/ is that the 32 ms frequency analysis window that is used as the first step in the creation of the spectrum envelope smears over a time interval that is too long to preserve the rapid increase in energy (or energy in a specific band) which probably characterizes /b/ but not /v/.[6] Inadequate time resolution is also consistent with the fact that all of the sounds in the top group of rows in Table II are characterized by rapid spectral changes. The middle group of rows, comparing the naturally spoken signals with their DSS counterparts, is also dominated by sounds whose recognition is known to depend on rapid spectral changes. There is clearly more involved than temporal resolution, however, since there are a few sounds showing sizeable drops in intelligibility that are not typically thought of in terms of rapid spectral movement (e.g., /ʒ/, /r/, /ʃ/, /h/) as well as many rapid-spectral-change sounds that do not appear on either the natural/SES or natural/DSS lists. These details aside, in relation to the primarily experimental goals of the study the most important finding from the consonant recognition tests is that the fine-detail-preserving SES method conveys only slightly more information about consonant identity than the coarse peaks-only DSS method.

IV. DISCUSSION

The primary goal of this study was to determine the effects on speech intelligibility of reconstructing speech using a purposely coarse method of estimating the spectrum envelope based entirely on the distribution of spectral peaks (DSS) versus a method that attempts to preserve the fine details of the envelope shape (SES). Taken in the broadest possible terms the results showed that nearly as much speech information is conveyed by the peaks-only DSS method as by the detail-preserving SES method. Just as clearly, however, every test showed some measurable advantage for spectral detail, although the differences were not large in absolute terms.

J. Acoust. Soc. Am., Vol. 119, No. 6, June 2006

Hillenbrand *et al.*: Spectral peaks versus spectral shape   4049

Sentence intelligibility for both the TIMIT and HINT signals was slightly (~1 percentage point) but reliably higher for the SES condition, and listeners requested significantly more repetitions of the DSS versions. Sentences from both databases were presented in quiet under good listening conditions. It is quite possible that the intelligibility advantage for the SES sentences would be greater under adverse listening conditions. In general, though, the fact that sentence intelligibility scores were so similar for the two methods tends to provide some support for cochlear implant processor strategies such as MPEAK (Multipeak) and SPEAK (Spectral Peak) that rely on transmitting primarily the high energy components of the spectrum. The results also suggest that accurate formant tracking is not necessary, at least for quiet signals, since the DSS is driven by unedited spectral peaks rather than edited formants. There is, however, evidence indicating that MPEAK, which relies on peak picking, does not perform as well in noise as SPEAK which, despite the name, transmits information about the highest amplitude spectral components, independent of whether the component constitutes a spectral peak (Skinner, Holden, Holden, Dowell, Seligman, Brimacombe, and Beiter, 1991).

The major findings from the vowel intelligibility tests are (1) there was no difference in vowel intelligibility between the SES signals and the naturally spoken signals, (2) there was an intelligibility drop of some 5–6 percentage points for both of peaks-only methods (DSS and formant synthesis), and (3) there was no difference in vowel intelligibility between the DSS and formant synthesis conditions. Taken together we believe these findings argue for an underlying pattern matching mechanism that is very much in line with Klatt's (1982) interpretation of his findings on phonetic distance judgments with static vowels. While Klatt's results showed that formant frequencies were by far the most important determinant of vowel quality, he argued that his findings were best explained not by an underlying formant tracking mechanism but rather by a spectral shape pattern matching process which, "… attends to the locations of prominent energy concentrations, but does not attend greatly to their relative intensities, nor to the shape of the spectrum in the valleys between energy concentrations" (p. 1281). Three aspects of the present vowel intelligibility results are consistent with this interpretation. First, a modest but measurable loss of phonetically relevant information occurs when the original spectral shape is reduced to either formants or unedited spectral peaks, indicating that formants/spectral peaks alone are not entirely sufficient to specify vowel identity [see Bladon's (1982), *reduction* argument in his critique of formant representations]. Second, formants *per se* do not appear to be required since vowel identity was conveyed just as well by unedited spectral peaks as by hand-edited formant frequencies. Third, the confusion matrix for the DSS vowels remains quite well behaved, showing almost exclusively listening errors involving vowels that are phonetically quite similar to the vowel intended by the talker. Given that the DSS is driven by raw spectral peaks, this suggests to us that this synthesis method conveys vowel identity as well as the formant synthesizer not by implicitly conveying formant information but rather by doing a reasonable enough job of pre-serving the shape of the spectrum in high energy regions.

There remains considerable uncertainty about the precise nature of the peak-dominated spectral shape pattern matching scheme that might meet Klatt's (1982) requirements. Klatt proposed a *weighted spectral slope* metric (WSM) based on the comparison of spectral slopes, but with greater weight given to slope differences in and around spectral peaks. The metric provided good predictions of perceived phonetic distances among static synthetic vowels. Klatt's WSM metric was used by Nocerino *et al.* (1985) in a study of automatic recognition of naturally spoken words (alphabet and digits). Surprisingly, WSM was found to perform best when the peak-emphasizing weighting factors were completely removed, a finding that runs counter to the notion that spectral shape differences in the vicinity of peaks are of greater importance. A similar result was reported in a study of concurrent vowel recognition by Assmann and Summerfield (1989), where the WSM showed only a small increase in recognition performance over a non-peak-weighted version of the same metric.

Hillenbrand and Houde (2003) proposed a *narrow band pattern matching model* of vowel perception based on city block spectral distances between narrow band input spectra and smoothed spectral shape templates derived empirically by averaging the harmonic spectra of like vowels spoken by a panel of talkers. The information-bearing spectral peaks were enhanced by a "thresholding" procedure that zeroes out spectral values below a threshold function consisting of a center-weighted running average of spectral amplitudes. The pattern matching model recognized vowels from a 1668-token database consisting of 12 vowels in /hVd/ context spoken by men, women, and children. Model accuracy approached but did not quite reach that of human listeners. Of special relevance to the present findings, the effect of the peak-enhancing operation was dramatic, with overall recognition accuracy falling from 91.4% with the thresholding operation all the way to 59.9% without it. Similarly, Liénard and Di Benedetto (2000) developed a peak-enhancement method to recognize French vowels from "bump vectors"— smooth spectra resulting from an independently developed peak-enhancing operation that is quite similar to the one used in our pattern matching model. Recognition experiments showed a substantial advantage for the bump vector over a variety of alternative smoothed spectral shape representations that did not incorporate a peak-enhancement operation.

The pattern of results for the consonant intelligibility tests differed in some important ways from the vowel results. There was a small but statistically reliable advantage of about 3 percentage points for the detail-preserving SES method as compared to the peaks-only DSS method. In numerical terms, however, the SES-DSS difference for consonants was only about half of that for vowels. Consonant spectra tend to be rather ragged and irregular in shape, and it appears that in most but not quite all cases the more coarse DSS method conveys the relevant phonetic information as well as the SES method. However, unlike the vowel tests, which showed that the SES signals were just as intelligible as the naturally spoken signals, the SES versions of the consonants were, on average, 8.5 percentage points less intelli-

Hillenbrand *et al.*: Spectral peaks versus spectral shape

gible than their naturally spoken counterparts. Especially large decrements in intelligibility tended to be seen for consonants that are cued by rapid changes in the spectrum, suggesting that inadequate temporal resolution was at least in part responsible for the intelligibility deficit. Follow up experiments are underway to explore this possibility.

Appendix A

TABLE A1. Confusion matrix for naturally spoken /hVd/ utterances (iy=/i/, ih=/ɪ/, ei=/e/, eh=/ɛ/, ae=/æ/, ah=/ɑ/, aw=/ɔ/, oa=/o/, oo=/ʊ/, uw=/u/, uh=/ʌ/, er=/ɝ/).

	iy	ih	ei	eh	ae	ah	aw	oa	oo	uw	uh	er
iy	97.5	0.9	0.9				0.3					0.3
ih	0.3	99.1		0.3					0.3			
ei		0.3	96.9	0.6	1.8					0.3		
eh				94.5	4.3		0.3			0.3	0.6	
ae				1.8	97.9	0.3						
ah				0.3	15.7	79.1	4.3		0.3		0.3	
aw	0.3				0.3	7.4	90.5	0.3			1.2	
oa				0.3			0.6	98.8	0.3			
oo			0.3				0.3		95.7	0.6	2.5	0.6
uw								2.2	3.7	94.1		
uh		0.3				1.2	0.6	0.3	1.5		95.7	0.3
er										0.3		99.7

TABLE A2. Confusion matrix for utterances generated with the spectral envelope synthesizer (iy=/i/, ih=/ɪ/, ei=/e/, eh=/ɛ/, ae=/æ/, ah=/ɑ/, aw=/ɔ/, oa=/o/, oo=/ʊ/, uw=/u/, uh=/ʌ/, er=/ɝ/).

	iy	ih	ei	eh	ae	ah	aw	oa	oo	uw	uh	er
iy	98.2		1.5	0.3								
ih	0.9	97.2		1.5					0.3			
ei		0.3	98.2	0.6		0.6				0.3		
eh			0.9	90.5	8.6							
ae				0.6	99.4							
ah			0.3		14.7	83.7	0.6	0.3	0.3			
aw						6.2	92.3				1.2	0.3
oa								99.4		0.6		
oo		0.3					0.3	0.3	93.9	0.3	4.0	0.9
uw								1.5	2.5	95.4	0.3	0.3
uh					0.3	1.5	1.2	0.3	1.8	0.6	93.9	0.3
er	0.3	0.3									0.3	99.1

TABLE A3. Confusion matrix for utterances generated with the damped sine wave synthesizer (iy=/i/, ih=/ɪ/, ei=/e/, eh=/ɛ/, ae=/æ/, ah=/ɑ/, aw=/ɔ/, oa=/o/, oo=/ʊ/, uw=/u/, uh=/ʌ/, er=/ɝ/).

	iy	ih	ei	eh	ae	ah	aw	oa	oo	uw	uh	er
iy	93.9	3.7	2.1	0.3								
ih	1.5	96.6		1.8								
ei	5.2	0.6	94.2									
eh			0.6	94.8	4.6							
ae		0.3	0.3	12.2	86.9					0.3		
ah					19.0	71.8	3.4	0.6			4.9	0.3
aw					0.3	15.5	79.3	0.6	0.6		3.7	
oa						0.3		86.2	3.4	5.2	4.9	
oo	0.3								95.1	2.5	1.8	0.3
uw	0.3							5.2	4.6	89.2	0.6	
uh			0.6			0.3	0.6		10.8		87.1	0.6
er			1.5		0.3	0.3		0.3	2.8	0.6	0.3	93.9

TABLE A4. Confusion matrix for utterances generated with the Klatt and Klatt (1990) formant synthesizer (iy=/i/, ih=/ɪ/, ei=/e/, eh=/ɛ/, ae=/æ/, ah =/ɑ/, aw=/ɔ/, oa=/o/, oo=/ʊ/, uw=/u/, uh=/ʌ/, er=/ɝ/).

	iy	ih	ei	eh	ae	ah	aw	oa	oo	uw	uh	er
iy	91.4	3.7	4.6			0.3						
ih	0.6	98.2		0.9		0.3						
ei	3.7	1.9	92.9	0.3	0.9	0.3						
eh		0.9		92.9	6.1							
ae		0.3		11.0	88.7							
ah				0.6	13.5	79.7	3.7	0.3			1.8	0.3
aw	0.3					4.3	95.1		0.3			
oa						0.6	0.6	93.9	2.5	1.5	0.6	0.3
oo			0.3	0.6					93.6	0.9	4.3	0.3
uw								0.3	18.1	14.1	67.5	
uh		0.3				0.9	1.5		5.8		91.4	
er	0.3								0.3			99.4

Appendix B

TABLE B1. Confusion matrix for the consonant intelligibility tests using naturally spoken utterances (th=/θ/, sh=/ʃ/, dh=/ð/, zh=/ʒ/, dz=/dʒ/,ch =/tʃ/). Percentages are rounded to the nearest integer.

	b	d	g	p	t	k	f	th	s	sh	h	v	dh	z	zh	dz	ch	m	n	l	r	w	j
b	97	1							1			1	1										
d		100																					
g		1	96			3																	
p				96		3																	
t					100																		
k						100																	
f							91	8	1														
th							3	93	2				2										
s								3	97														
sh									1	99							1						
h				2							95											3	
v												99	2										
dh		1							4			1	93										
z														99									
zh															97	3							
dz		1	5													94	1						
ch					1					2							97						
m																		100					
n																			99				
l	2																			98			
r	1		1																		98		
w												1										98	1
j																							100

	b	d	g	p	t	k	f	th	s	sh	h	v	dh	z	zh	dz	ch	m	n	l	r	w	j
b	41	1						1				50	5					2					
d		81	1					1					15			1							
g		2	87	7		2						1											1
p				93			1				6												
t					92											1	7						
k			14	1		81					3						1						
f							92	8	1			1											
th							9	78	2			1	10										
s								2	98														
sh									1	97				1			1						
h				6			1	1			89		1									3	
v								1				99	1										
dh								3				11	86										
z												1	6	94									
zh															89	7							4
dz			6												5	88	2						
ch										16							83						
m																		100					
n						1												3	96				
l	5												2							92			
r	8		1																		92		
w												4										96	
j	1																			1			98

	b	d	g	p	t	k	f	th	s	sh	h	v	dh	z	zh	dz	ch	m	n	l	r	w	j
b	74	2	3					1				13	5										1
d		81	2		1			1					13			1							
g		7	84	1		3		1				1	1										2
p				88		7					4							1					
t					97	1											1						
k				10	1	85					3												
f							62	36			1	1	1										
th							6	86	1		1		6										
s								1	98	1													
sh							1		6	84	1				1		6						
h				11		1		1			84			1								3	
v	4											78	10					4		5	1		
dh		1						3				2	91					3					
z													5	94									
zh			1									7	3	74	8					1			5
dz			2	5			1							1	5	79	7					1	
ch					3				1	16						1	79						
m	1											2						95		1			1
n																		7	92	1			
l	4			1									2					1		90			
r	7		1															4		3	81	4	
w												2						2				95	1
j			1															5		1			93

[1]The most important difference between the implementation used here and that described in Hillenbrand and Houde (2002) involves the FFT window size for envelope analysis, which was shortened from 64 to 32 ms, based on findings showing poor identification of some phonetic contrasts that depend on rapid spectral change. As will be discussed below, time resolution remains a problem even with the 32 ms window.

[2]Throughout the text, we will use the term *Gaussian-weighted running average* to refer to an approximation implemented with three passes of a rectangular (i.e., unweighted) running average. In a Gaussian weighted running average, each spectral amplitude is replaced by the weighted average of n neighbors of higher and lower frequency, with 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 (i.e., unweighted) average: 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. (See Hillenbrand and Houde (2003), for an explanation of the end-correction scheme that is used.).

[3]Without a smoothing operation of some kind, the peak frequencies that are measured from the harmonic envelope of a voiced speech segment would always correspond to the highest amplitude harmonic in the collection of harmonics comprising the peak. The evidence is quite clear that the phonetic quality associated with speech signals does not correspond to the strongest harmonic (Klatt, 1986). A peak frequency measured from the harmonic envelope after F_0-dependent smoothing corresponds closely with the method based on the weighted average of harmonic amplitudes used in studies such as Peterson and Barney (1952); i.e., for a peak composed of three harmonics, the peak frequency will be displaced from the central harmonic in the direction of the second strongest harmonic.

[4]What is referred to here as a "thresholding" operation is referred to in some of our earlier writings as a masking operation, a term which remains appropriate, in our view, but was dropped to avoid creating the impression that we were attempting a faithful simulation of simultaneous masking in the auditory system.

[5]The weak fricative /θ/, which was confused primarily with /f/, might seem out of place with the stops and the affricate which form the remainder of this list, but there is clear evidence that the /f/-/θ/ distinction is cued primarily by brief-duration formant transitions rather than differences in the spectral content of the quasistationary fricative noise (Harris, 1958).

[6]The word *probably* is used here because we are unaware of any published experimental work that has systematically examined the cues to distinctions between stops and similar-place weak fricatives such as /b/-/v/ and /d/-/ð/. Extensive pilot work of our own suggests that the voiced stops /b/ and /d/ typically show a more abrupt increase in high-frequency energy (e.g., above 1 kHz) than /v/ and /ð/. Further, because of the 32 ms analysis window size, this abrupt high-frequency onset that appears to characterize stops is often not observed in the SES versions of /b/ and /d/. Informal listening also suggests to us that SES versions of /b/ are rendered much more faithfully with short analysis window sizes such as 8 ms, although other features of the synthesis are severely compromised when this short window is used.

Assmann, P. F., and Summerfield, Q. (**1989**). "Modeling the perception of concurrent vowels: Vowels with the same fundamental frequency," , **85**, 327–338.

Bench, J., Kowal, A., and Bamford, J. (**1979**). "The BKB (Bamford-Kowal-Bench) sentence lists for partially-hearing children," Br. J. Audiol., **13**, 108–112.

Bladon, A. (**1982**). "Arguments against formants in the auditory representation of speech," in *The Representation of Speech in the Peripheral Auditory System*, edited by R. Carlson and B. Granstrom (Elsevier Biomedical Press, Amsterdam), pp. 95–102.

Bladon, A. and Lindblom, B. (**1981**). "Modeling the judgment of vowel quality differences," J. Acoust. Soc. Am., **69**, 1414–1422.

Carlson, R., and Granstrom, B. (**1979**). "Model predictions of vowel dissimilarity," Speech Transmission Laboratory Quarterly Progress and Status Report No. STL-QPSR 3-4/1979 (Royal Institute of Technology, Stockholm, Sweden), pp. 84–104.

Carlson, R., Granstrom, B., and Klatt, D. H. (**1979**). "Vowel perception: The relative perceptual salience of selected acoustic manipulations," Speech Transmission Laboratory Quarterly Progress and Status Report No. STL-QPSR 3-4/1979 (Royal Institute of Technology, Stockholm, Sweden), pp. 73–83.

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G.,Pallett, D. S., and Dahlgren, N. L. (**1993**). "DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM," Gaithersburg, MD, National Institute of Standards and Techonology.

Harris, K. S. (**1958**). "Cues for the discrimination of American English fricatives in spoken syllables," Lang Speech, **1**, pp. 1–7.

Hillenbrand, J. M., and Gayvert, R. T. (**2005**). "Open source software for experiment design and control," J. Speech Lang. Hear. Res., **48**, 45–60.

Hillenbrand, J., Getty, L. A., Clark, M. J., and Wheeler, K. (**1995**). "Acoustic characteristics of American English vowels," J. Acoust. Soc. Am., **97**, 1300–1313.

Hillenbrand, J. M., and Houde, R. A. (**1996**). "A method for creating filters with arbitrary response characteristics for use in hearing and speech research," J. Speech Hear. Res., **39**, 390–395.

Hillenbrand, J. M., and Houde, R. A. (**2002**). "Speech synthesis using damped sinusoids," J. Speech Lang. Hear. Res., **45**, 639–650.

Hillenbrand, J. M., and Houde, R. A. (**2003**). "A narrow band pattern-matching model of vowel perception," J. Acoust. Soc. Am., **113**, 1044–1055.

Hillenbrand, J. M., and Nearey, T. N. (**1999**). "Identification of resynthesized /hVd/ syllables: Effects of formant contour," J. Acoust. Soc. Am., **105**, 3509–3523.

Klatt, D. (**1980**). "Software for a cascade/parallel formant synthesizer," J. Acoust. Soc. Am., **67**, 971–995.

Klatt, D. H. (**1982**). "Prediction of perceived phonetic distance from critical-band spectra: A first step," *IEEE ICASSP*, 1278–1281.

Klatt, D. H. (**1986**). "Representation of the first formant in speech recognition and in models of the auditory periphery," in *Proceedings of the Montreal Satellite Symposium on Speech Recognition*, edited by P. Mermelstein, McGill University, Montreal, pp. 5–7.

Klatt, D. H., and Klatt, L. C. (**1990**). "Analysis, synthesis, and perception of voice quality variations among female and male talkers," J. Acoust. Soc. Am., **87**, 820–857.

Liénard, J-S., and Di Benedetto, M-G. (**2000**). "Extracting vowel characteristics from smoothed spectra," J. Acoust. Soc. Am., **108**, Suppl. 1, 2602.

Lindblom, B. (**1978**). "Phonetic aspects of linguistic explanation," Studia Linguistica, **32**, 137–153.

Nilsson, M., Soli, S. D., and Sullivan, J. A. (**1994**). "Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise," J. Acoust. Soc. Am., **95**, 1085–1099.

Nocerino, N., Soong, F. K., Rabiner, L. R. and Klatt, D. H., (**1985**). "Comparative study of several distortion measures for speech recognition," Speech Commun., **4**, 317–331.

Paul, D. B. (**1981**). "The spectral envelope estimation vocoder," IEEE Trans. Acoust., Speech, Signal Process., **29**, 786–794.

Peterson, G., and Barney, H. L. (**1952**). "Control methods used in a study of the vowels," J. Acoust. Soc. Am., **24**, 175–184.

Shannon, R. V., Jensvold, A. Padilla, M., Robert, M., and Wang, X. (**1999**). "Consonant recordings for speech testing," J. Acoust. Soc. Am., **106**, 71–74.

Skinner, M., Holden, L., Holden, T., Dowell, R., Seligman, P., Brimacombe, J., and Beiter, A. (**1991**). "Performance of postlinguistically deaf adults with the Wearable Speech Processor (WSP III) and Mini Speech Processor (MSP) of the Nucleus multi-electrode cochlear implant," Ear Hear., **12**, 3–22.

Zahorian, S. A., and Jagharghi, A. J. (**1993**). "Spectral shape versus formants as acoustic correlates for vowels," J. Acoust. Soc. Am., **94**, 1966–1982.