A narrow band pattern-matching model of vowel perception

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

Robert A. Houde
RIT Research Corporation, 125 Tech Park Drive, Rochester, New York 14623

(Received 5 February 2002; accepted for publication 2 August 2002)

The purpose of this paper is to propose and evaluate a new model of vowel perception which assumes that vowel identity is recognized by a template-matching process involving the comparison of narrow band input spectra with a set of smoothed spectral-shape templates that are learned through ordinary exposure to speech. In the present simulation of this process, the input spectra are computed over a sufficiently long window to resolve individual harmonics of voiced speech. Prior to template creation and pattern matching, the narrow band spectra are amplitude equalized by a spectrum-level normalization process, and the information-bearing spectral peaks are enhanced by a ''flooring'' procedure that zeroes out spectral values below a threshold function consisting of a center-weighted running average of spectral amplitudes. Templates for each vowel category are created simply by averaging the narrow band spectra of like vowels spoken by a panel of talkers. In the present implementation, separate templates are used for men, women, and children. The pattern matching is implemented with a simple city-block distance measure given by the sum of the channel-by-channel differences between the narrow band input spectrum (level-equalized and floored) and each vowel template. Spectral movement is taken into account by computing the distance measure at several points throughout the course of the vowel. The input spectrum is assigned to the vowel template that results in the smallest difference accumulated over the sequence of spectral slices. The model was evaluated using a large database consisting of 12 vowels in /hVd/ context spoken by 45 men, 48 women, and 46 children. The narrow band model classified vowels in this database with a degree of accuracy (91.4%) approaching that of human listeners. [DOI: 10.1121/1.1513647]

PACS numbers: 43.71.An, 43.72.Ar, 43.66.Ba [KRK]

I. INTRODUCTION

A longstanding goal of speech perception research is to explain the perceptual mechanisms that are involved in the recognition of vowel identity. As with most other phenomena in phonetic perception, this problem has resisted straightforward solution, due in large measure to the variation in acoustic patterns that is observed when like vowels are spoken by different talkers, in different phonetic environments, at different speaking rates, at different fundamental frequencies, or with varying levels of contrastive stress. A wide range of models have been proposed to address one or more of these variability problems (for a review, see Rosner and Pickering, 1994). This diversity in theoretical approaches stands in contrast to a rather limited set of choices in the underlying acoustic representations that drive these recognition models. The overwhelming majority of vowel identification models have assumed that the recognition process is driven by an underlying representation consisting of either the formant frequency pattern of the vowel (with or without fundamental frequency as a normalizing factor) or the gross shape of the smoothed spectral envelope. Competition between these two quite different approaches to vowel identification has occupied a fair amount of attention in the literature. Excellent reviews of this literature can be found in Bladon and Lindblom (1981), Bladon (1982), Klatt (1982a, b), Zahorian and Jagharghi (1986), and Ito *et al.* (2001). To summarize the issues briefly, the main idea underlying formant representations is the notion that the recognition of vowel identity (and many other aspects of phonetic quality) is controlled not by the detailed shape of the spectrum but rather by the distribution of formant frequencies, chiefly the three lowest formants (F_1–F_3). The virtues of formant representations include the following: (1) formant representations are quite compact relative to whole spectrum models, in keeping with commonly held notions about the low dimensionality of perceptual space for vowels (e.g., Pols *et al.*, 1969); (2) formant representations allow for a number of fairly straightforward solutions to talker normalization problems that arise from variation in vocal tract length (e.g., Disner, 1980; Miller, 1989; Nearey, 1992; Nearey *et al.*, 1979; Syrdal and Gopal, 1986; Hillenbrand and Gayvert, 1993); (3) reasonable correspondences have been found between formant representations and perceptual dimensions inferred from measures of perceived similarities among vowels (e.g., Miller, 1956; Pols *et al.*, 1969; cf. Bladon and Lindblom, 1981); (4) speech or speechlike signals that are synthesized from formant representations are typically highly intelligible, even in cases involving gross departures in detailed spectral shape between the original and reconstructed signals (e.g., Remez *et al.*, 1981); (5) several pattern recognition studies have shown that naturally spoken vowels can be recognized with reason-

a)Electronic mail: james.hillenbrand@wmich.edu

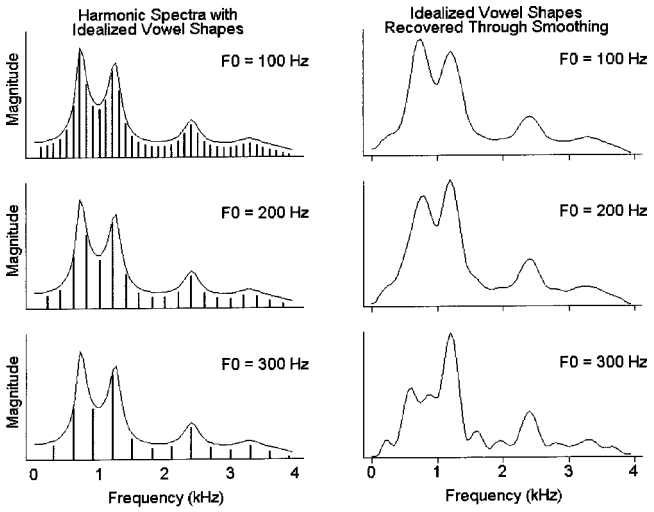

FIG. 1. Left: Harmonic spectra and estimates of the idealized smooth envelope shapes (smooth curves—see text) for the vowel /ɑ/ spoken at three fundamental frequencies, but with the same vocal tract frequency response curve and glottal source shape. Right: Attempting to recover the idealized envelope shape through cepstral smoothing. Note that the smoothing operation recovers the vocal tract filter shape well at 100 Hz, reasonably well at 200 Hz, but very poorly at 300 Hz. The distortion in the smoothed envelope shape at $F_0 = 300$ Hz is due to aliasing. For a detailed discussion of this phenomenon, see de Cheveigné and Kawahara (1999).

able accuracy based on formant measurements, particularly when fundamental frequency, formant movements, and duration are taken into account (e.g., Nearey *et al.*, 1979; Syrdal and Gopal, 1986; Hillenbrand and Gayvert, 1993; Hillenbrand *et al.*, 1995, 2000b); and (6) probably of greatest importance, the widely cited evidence from Klatt (1982a) showing that formant frequencies are easily the most important acoustic parameters affecting vowel quality, with other manipulations of spectral shape (e.g., high- and low-pass filtering, spectral tilt, formant amplitudes, notches, formant bandwidths, etc.) resulting in little or no change in phonetic quality (see also related evidence from Assmann and Summerfield, 1989). There are, however, quite substantial problems with formant theory (see Bladon *et al.*, 1982; Zahorian and Jagharghi, 1993; Ito *et al.*, 2001 for reviews). These problems include (1) the unresolved and quite possibly unresolvable problem of tracking formants in natural speech; (2) evidence showing that some spectral details other than formant frequencies can affect vowel quality (e.g., Chistovich and Lublinskaya, 1979; Bladon, 1982); and (3) the observation that errors made by human listeners do not appear to be consistent with the idea that vowel quality is perceived on the basis of labeled formants (Klatt, 1982b; Ito *et al.*, 2001). The primary weakness of spectral shape approaches is generally thought to be the difficulty of this class of models to accommodate what would appear to be rather compelling evidence showing that large changes can be made in detailed spectral shape without affecting phonetic quality, as long as the formant frequency pattern is preserved (e.g., Remez *et al.*, 1981; Klatt, 1982a).

A feature that is common to nearly all spectral shape models is the derivation of the spectral envelope through some kind of smoothing operation. The motivation underlying the smoothing is straightforward. The left panel of Fig. 1 shows the vowel /ɑ/ spoken at three fundamental frequencies, but with the same vocal tract frequency response curve and glottal source shape. Vowel quality will be similar in the three cases (but not identical—e.g., Miller, 1953; Fujisaki and Kawashima, 1968) despite the obvious differences in their harmonic spectra. Smoothing is intended to remove the largely irrelevant harmonic detail. The smooth curves to the left in this figure represent the idealized smooth shape for this vowel. This idealized shape corresponds to a hypothetical filter representing the combined effects of the vocal tract frequency response curve, the shape of the glottal source spectrum, and the radiation characteristic. (Hereafter, to avoid this awkwardly long phrase, we will refer simply to the *vocal tract filter*, although it should be understood that the detailed shape of this hypothetical equivalent filter is controlled by the vocal tract transfer function, the spectrum of the source signal, and the radiation characteristic. Easily the most important of these three functions in signaling differences in vowel identity is the vocal tract frequency response curve.) As de Cheveigné and Kawahara (1999) note, for the case of phonated vowels, the vocal tract filter function is effectively sampled at discrete frequencies corresponding to the harmonics of the voice source. As such, the ability to estimate the shape of the vocal tract filter based on the amplitudes of the voice-source harmonics is subject to the well-known constraints of sampling theorem. de Cheveigné and Kawahara describe a series of simple tests that clearly demonstrate that the vocal tract transfer function is severely undersampled at higher fundamental frequencies. Of direct relevance to the present discussion, de Cheveigné and Kawahara go on to show that the smoothing that is so commonly used in spectral shape models results in aliasing-induced distortion of the estimated vocal tract filter function at even moderately high fundamental frequencies (see especially Fig. 4 of de Cheveigné and Kawahara, which provides a very concise summary of the aliasing problem). This aliasing effect is illustrated in the right half of Fig. 1, which shows smoothed spectra derived by a cepstral smoothing operation. Note that the smoothing operation recovers the vocal tract filter shape well at 100 Hz, reasonably well at 200 Hz, but very poorly at 300 Hz.

In response to this undersampling problem, de Cheveigné and Kawahara proposed a novel *missing data model* of vowel identification in which the *unsmoothed* harmonic spectrum is directly compared to a set of smoothed templates, with spectral differences computed *only at frequencies corresponding to voice-source harmonics*. Small-scale simulations of two versions of the *missing data model* showed good recognition of five synthetic vowels generated at a wide range of fundamental frequencies (20–300 Hz) when compared against templates consisting of known spectral envelopes for each vowel. An otherwise identical model using smoothed input spectra showed extreme sensitivity to fundamental frequency.

The purpose of the present article is to address two important limitations of the innovative work described by de Cheveigné and Kawahara (1999), one theoretical, the other experimental. A central feature of the *missing data model* is the notion that the input spectrum is compared to the tem-

plates only at harmonic frequencies. This is a reasonable restriction, for exactly the reasons discussed by de Cheveigné and Kawahara, but it is one which imposes some rather strict signal processing demands on the pattern recognizer. The instantaneous fundamental frequency must be estimated, presumably with considerable precision since it would not take much of a pitch estimation error to result in a situation in which the pattern recognizer ended up comparing the input spectrum with the template exclusively at frequencies that are irrelevant to vowel identity. Further, as the authors note, modifications would have to be made to the *missing data model* to handle aperiodic or marginally periodic signals, such as whispered or breathy vowels or vowels spoken with rapid changes in the fundamental frequency. de Cheveigné and Kawahara offer some reasonable speculations about methods that might be used to address these kinds of cases, but in the present article we will propose a vowel identification model which removes the harmonics-only restriction entirely. In common with the *missing data model*, our *narrow band pattern matching model* compares unsmoothed, high-resolution input spectra (i.e., spectra computed over a sufficiently long window to resolve voice-source harmonics) directly with a set of smooth vowel templates. However, unlike the *missing data model*, which computes spectral distances only at harmonic frequencies, spectral distances in our *narrow band model* are computed at all frequencies. Results will be presented which we believe demonstrate that the harmonics-only restriction needlessly complicates the pattern matching and that a recognition algorithm that gives no special treatment to harmonics accurately recognizes vowels spoken with a wide range of fundamental frequencies.

The second limitation which we wish to address is that the *missing data model* was not evaluated on naturally spoken utterances. The test signals that were classified by the model consisted of synthetic vowels generated with perfectly periodic source signals and static spectral envelopes. Further, the templates for the five vowel types that were used consisted of the known transfer functions that were used in generating the test signals. While these tests were quite reasonable in light of the goals of de Cheveigné and Kawahara's paper (demonstrating the aliasing effects discussed above), it remains unclear how well a model that is driven by unsmoothed harmonic spectra would classify natural spoken vowels. In the present work, our narrow band model will be tested using naturally spoken utterances comprising 12 American English vowels produced by a large group of men, women, and children. Further, the vowel templates will be derived empirically based on an analysis of those naturally spoken utterances.

II. THE NARROW BAND PATTERN-MATCHING MODEL

A. Preliminary comments

Before describing the details of the narrow band model, we should note that the model does not represent an attempt at a faithful simulation of the physiological response properties of the auditory system. Indeed, we believe that the clearest lesson of Klatt's (1982a) study is that a model of peripheral auditory processing in and of itself is inherently incapable of accounting for the recognition of vowel quality. Klatt's findings show that there are many aspects of spectral shape that are quite audible (and therefore must be preserved in any faithful auditory model) but which contribute very little to judgments of vowel timbre. A complete model of vowel recognition would need to incorporate not only a precise simulation of the low-level auditory representations from which generic judgments of timbre might be derived, but also of the (presumably) decision-level psychological mechanisms that must be involved in translating these low-level auditory representations into vowel percepts (i.e., ignoring, or largely ignoring, features such spectral tilt, formant amplitude relations, spectral notches, etc.). In the present work we have adopted a simpler approach in which a few signal processing steps are intended to model the most psychologically important aspects of vowel recognition, with both low-level analysis and decision-level mechanisms rolled into single process. In general, our strategy in constructing the model was primarily one of working backward from key perceptual data such as those of Klatt (1982a) and Ito *et al.* (2001) toward a psychologically plausible processing scheme rather than one of working forward from peripheral auditory processing principles.

B. Template creation

Our model assumes that the reference patterns defining each vowel category consist of *sequences* of smooth spectra. We assume a sequence of spectra rather than a single spectrum sampled at steady state because of the large body of evidence implicating a strong role for spectral movement in vowel recognition (e.g., Strange *et al.*, 1983; Nearey and Assmann, 1986; Jenkins *et al.*, 1983; Parker and Diehl, 1984; Andruski and Nearey, 1992; Jenkins and Strange, 1999; Hillenbrand and Gayvert, 1993; Hillenbrand and Nearey, 1999; Assmann and Katz, 2000, 2001). We assume that the individual spectral shape templates in the sequence are derived simply by averaging the narrow band spectra of like vowels sampled at comparable time points throughout the course of the vowel (followed by light smoothing—see below). Figure 2 shows a sequence of templates for /æ/ sampled at 15%, 30%, 45%, 60%, and 75% of vowel duration derived from adult male talkers using tokens from a large /hVd/ database (described below). Figure 3 shows the signal processing steps that are used to generate the individual spectra that are averaged to produce smooth templates such as those illustrated in Fig. 2. The signal processing steps, which are detailed below, consisted of (1) calculation of a narrow band (i.e., long time window) Fourier spectrum, (2) spectrum-level normalization, (3) enhancement of spectral peaks by a thresholding procedure, and (4) overall amplitude normalization. The initial Fourier spectrum was computed over a relatively long time window. In the experiments reported below, we used a 512-point (64 ms) Hamming-windowed FFT with 8 kHz sampled waveforms and 6 dB per octave high-frequency preemphasis. Linear frequency and amplitude scales were used [see panel (a) of Fig. 3]. The next step involved the application of a broadband spectrum-level normalization (SLN) function. The motivation behind this

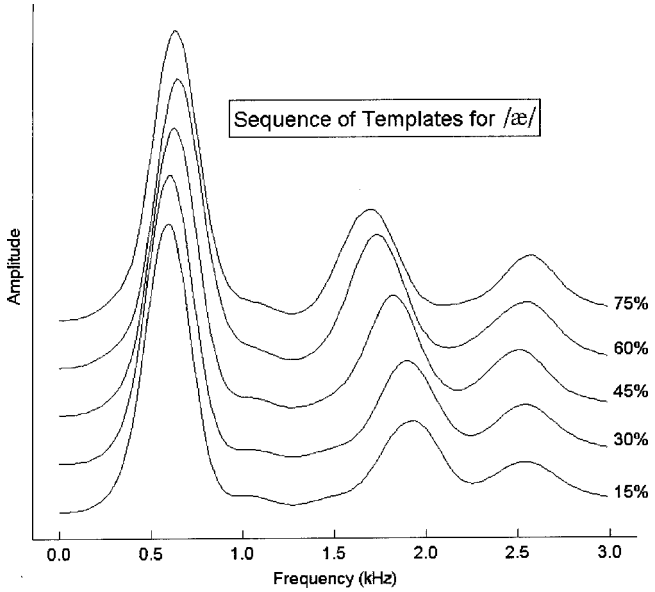

FIG. 2. Sequence of five vowel templates for /æ/ computed at 15%, 30%, 45%, 60%, and 75% of vowel duration. Note that successive templates have been offset on the amplitude scale so that the change in spectral shape over time can be seen more clearly.

step was to reduce as much as possible within-vowel-category differences in formant amplitude relations, following data such as Klatt (1982a) indicating that formant-amplitude variation, while quite audible to listeners, contributes little to perceived vowel color. The idea of the SLN operation, then, was simply to attenuate spectral regions of relatively high amplitude and amplify regions of relatively low amplitude, reducing the magnitude of amplitude differences among broad spectral peaks. The SLN operation was implemented by computing a gain function that

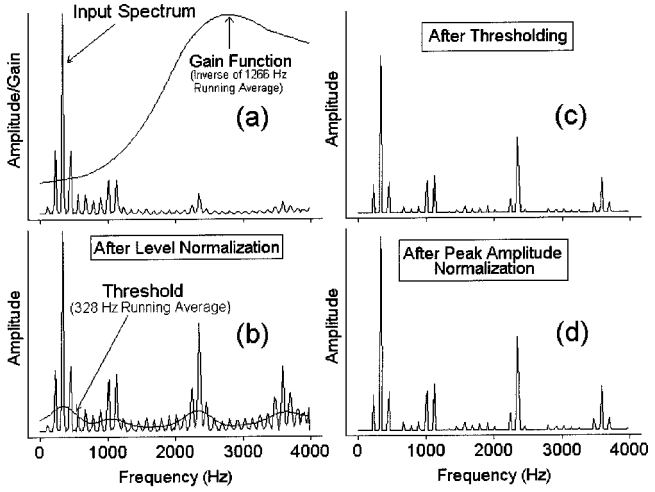

FIG. 3. Signal processing steps used in the narrow band pattern matching model: (a) FFT computed over a 64-ms Hamming-windowed segment and broadband spectrum-level normalization (SLN) function computed as the inverse of a 1266-Hz Gaussian-weighted running average of spectral amplitudes; (b) spectrum after processing by the SLN operation and a threshold function computed as a 328-Hz Gaussian-weighted running average of spectral amplitudes; (c) spectrum after thresholding, i.e., after zeroing all spectral values that lie below the threshold function; and (d) amplitude normalization, implemented by scaling the largest peak in the spectrum to a constant.

was relatively low in spectral regions with high average amplitude and, conversely, was relatively high in spectral regions with low average amplitude. The SLN function that is shown in panel (a) of Fig. 3 is simply the inverse of the Gaussian-weighted running average[1] of spectral amplitudes computed over an 81-channel (1265.6 Hz) spectral window. Panel (b) of Fig. 3 shows the spectrum after application of the broadband SLN operation. It can be seen that the variation in spectral peak amplitudes has been considerably reduced, although by no means entirely eliminated. The size of the smoothing window is a compromise, determined by inspecting a large number of individual normalized and unnormalized spectra. The rather large window size that was selected represents a compromise between two competing considerations. Very large window sizes produce rather limited benefit with respect to the goal of minimizing the importance of formant amplitude differences but have the advantage that they seldom amplify minor spectral peaks that are of little or no perceptual relevance. Smaller window sizes, on the other hand, do an excellent job of reducing the range of formant amplitude variation but can sometimes have the undesirable effect of amplifying minor spectral peaks. All of the informal experimentation involved in selecting a smoothing window size was carried out using a CVC database (Hillenbrand *et al.*, 2000b) other than the one used to evaluate the model.

The next signal processing step consisted of a thresholding procedure. The idea here was simply to emphasize the spectral peak regions (both narrow and broad band) that are known to have the greatest influence on vowel identity and to suppress the largely irrelevant spectral components in between harmonic peaks and in the less perceptually significant valleys that lie in between broad spectral peaks. This step was implemented by defining a threshold function as the Gaussian-weighted running average of spectral amplitudes computed over a 21-channel (328.1 Hz) spectral window. The running average is then subtracted from the spectrum, with all negative values (i.e., values below the threshold) set to zero.[2] As with the gain function described above, the size of the averaging window used for the threshold operation was determined through extensive informal experimentation using a vowel database other than the one used to evaluate the model. The process involved examination of a large number of individual cases of spectra with and without the thresholding operation and trying to find a smoothing window size that appeared to do the best job of enhancing the information-bearing aspects of the spectra (i.e., harmonics, especially those defining formant peaks). The final signal processing step involved amplitude normalization, implemented by scaling the largest peak in the spectrum to a constant [Fig. 3(d)].

In the tests reported below, separate template sequences were created for men, women, and children. These templates were created by averaging like vowels at like time points throughout the course of the vowel. For most of the tests reported below, we represent each vowel as a sequence of five templates, centered at 15%, 30%, 45%, 60%, and 75% of vowel duration, based on hand-measured values of vowel start and stop times from Hillenbrand *et al.* (1995). (The is-

sue of how many spectral slices are required will be considered below.) For example, the first spectrum of the /i/ template sequence for men was created by averaging all adult male tokens of /i/ sampled at 15% of vowel duration. Following the averaging, we apply light smoothing to each of the averaged spectra with a 171.9-Hz Gaussian-weighted running average. We assume that this final smoothing step has no counterpart in human perception and is a simple concession to the fact that we were forced by practicalities to create the templates from a limited number of tokens of each vowel. As will be explained below, each template was constructed by averaging roughly 40 examples of each vowel for each talker group. Although this is a relatively large database by experimental standards, it is a small fraction of the amount of speech a listener is likely to hear during the course of even a single day.

We attach no special importance to the choice of a sequence of five spectra to represent each vowel, as opposed to three or seven or any number of other reasonable alternatives. We will, however, report results demonstrating the simple point, in keeping with a good deal of evidence from human listeners, that a template sequence produces far better recognition performance than a pattern-matching scheme based on templates computed at a single time slice. Similarly, the choice of equally spaced samples between 15% and 75% is somewhat arbitrary. For illustration, Fig. 4 shows templates sampled at 30% of vowel duration for men (panel a), women (panel b), and children (panel c). Note that the formant frequency patterns that are traditionally associated with these vowels are well preserved in most but not all of these templates. For example, note the merger of F_1 and F_2 in the adult female /ɔ/ template (F_2 does not quite merge with F_1 in the child /ɔ/ template, but instead shows up as a soft shoulder rather than a peak) and the merger of F_2 and F_3 in all three /ɝ/ templates.

C. Computing distances between input spectra and template sequences

The distance between the input spectrum and a smoothed template for any given time slice is computed as the sum of the channel-by-channel absolute differences between the two spectra, divided by the total energy in the template, computed as the sum of template amplitudes across all 256 channels (i.e., a city-block distance, normalized for overall template amplitude). Unlike the method described by de Cheveigné and Kawahara (1999), which computes distances at voice-source harmonics only, distances are computed for all channels. Spectral differences between the narrow band input spectra and the smooth templates will obviously be large at frequencies remote from the harmonics, especially at high F_0, and especially in the deep valleys between the harmonics. A key assumption underlying our model is that these interharmonic differences should be approximately equally large to all vowel templates, presumably resulting in a more-or-less constant source of noise across all templates.

At each individual time slice, the spectral-distance algorithm produces a 12-element vector containing distances between the input spectrum and templates for each of the 12 vowel types derived from that same time slice (see Fig. 5). For illustration, Table I shows a set of distance vectors for a single token of /hæd/ using a method incorporating distances sampled at five time slices (15%, 30%, 45%, 60%, and 75% of vowel duration). The final distance vector used for recognition is created from the five individual vectors simply by computing a weighted average of the distances computed at each of the five time points. The recognition algorithm chooses the vowel corresponding to the smallest token-to-template distance. A weighted average was used based on an expectation that recognition performance would be improved if somewhat more weight were assigned to distances computed early in the vowel (typically corresponding to ''steady-state'' times measured in studies such as Peterson and Bar-

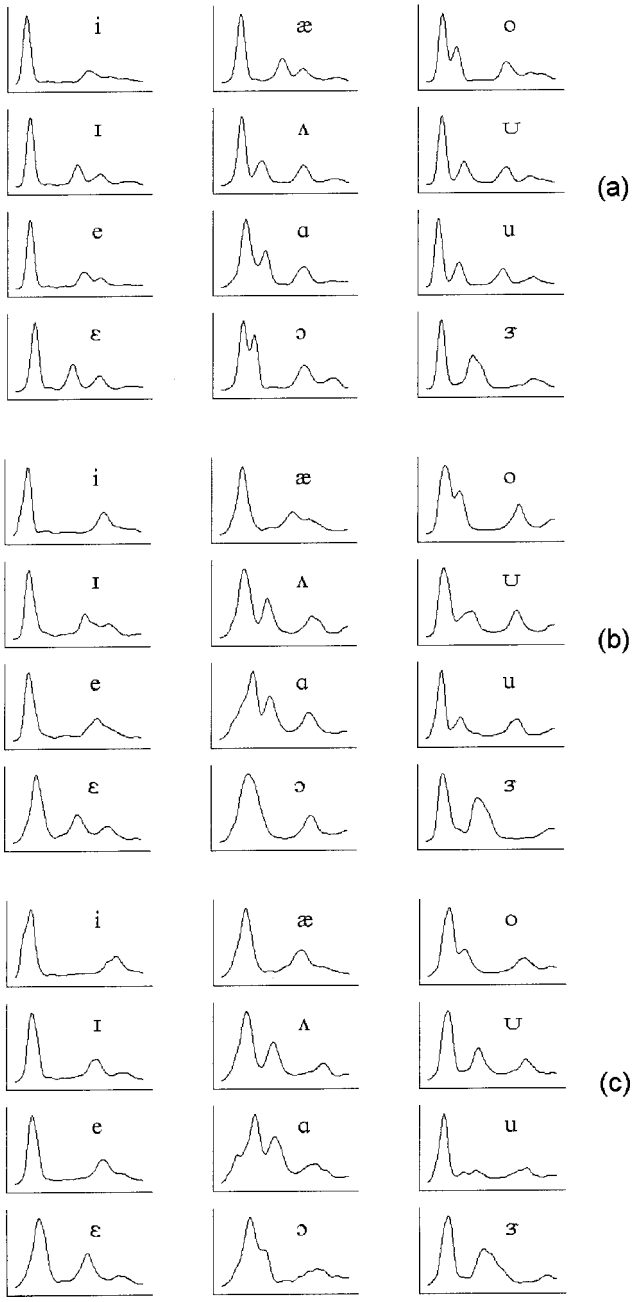

FIG. 4. Vowel templates computed at 30% of vowel duration for men (panel a), women (panel b), and children (panel c). The frequency scales are linear with a range of 0 to 4 kHz.

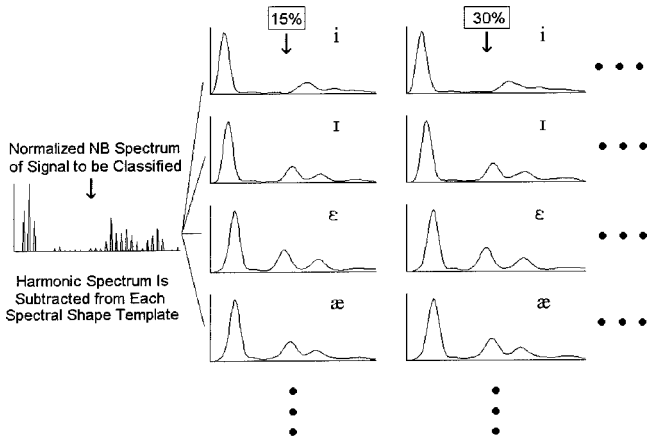

FIG. 5. Illustration of the recognition algorithm used in the narrow band model. The narrow band spectrum computed at 15% of vowel duration is compared to the 12 vowel templates computed at the same time point (only 4 of which are shown here); the narrow band spectrum at 30% of vowel duration (not shown) is then compared to the 12 vowel templates computed at 30% of vowel duration, and so on.

ney, 1952) than to offglide portions of the vowel which show the influence of the final consonant. Testing with a wide variety of weighting schemes showed that this was, indeed, the case. However, the advantage of a weighted over an unweighted average proved to be very slight. The five-slice results reported below used a scheme that assigned a weight of 1.0 to the first four slices and a weight of 0.6 to the last slice. The minimum distance in each row of Table I is shown in bold. Note that for the first three time slices, the minimum distance is to the (incorrect) /ɛ/ template. However, the minimum distance in the weighted-average vector, which takes the offglide into account as well, corresponds to the /æ/ category that was intended by the talker.

III. EVALUATION

The test signals consisted of 1668 naturally spoken /hVd/ utterances recorded by Hillenbrand *et al.* (1995). This database consists of 12 vowels (/i,ɪ,e,ɛ,æ,ɑ,ɔ,o, U,u,ʌ,ɝ/) spoken by 45 men, 48 women, and 46 10- to 12-year-old children. The stimuli, which were originally digitized at 16 kHz, were digitally low-pass filtered at 3.7 kHz, down-sampled to 8 kHz, and scaled to maximum peak amplitude. Omitted from the database *for the purposes of template creation only* were 192 utterances which had shown identification error rates of 15% or greater in the listening study described in the original 1995 article. The 1476-syllable stimulus set used for template construction consisted of roughly equal numbers of tokens spoken by men (482), women (522), and children (472). The full database of 1668 utterances was used during the recognition phase of the evaluation.

IV. RESULTS

Overall classification accuracy for the five-slice method was 92.0% for men, 92.2% for women, and 90.0% for children, with a mean 91.4% across the three talker groups. The corresponding intelligibility figures for human listeners, as measured in Hillenbrand *et al.* (1995), are 94.6%, 95.6%, and 93.7%, with a mean 94.6% across the three groups. Although the accuracy of the narrow band recognizer was ~3% poorer than the listeners, it is important to note that the listeners had access to duration cues, and there is clear evidence that duration has a modest but significant influence on vowel identification by human listeners. For example, Hillenbrand *et al.* (2000a), using resynthesized versions of a 300-utterance subset of the /hVd/ signals tested here, found a 2% decrease in vowel intelligibility when vowel duration was eliminated as a cue by setting the duration of all vowels to a neutral value. Further, overall vowel intelligibility was reduced by 4.5% to 5.0% when vowels were either shortened (fixed at 144 ms, 2 standard deviations below the grand mean of all vowel durations) or lengthened (fixed at 400 ms, 2 standard deviations above the grand mean). The current version of the narrow band model uses spectral information only.

Table II shows a confusion matrix relating the vowel intended by the talker to the vowel recognized by the model. The matrix combines results from all three talker groups. The overall look of the confusion matrix is similar in many respects to comparable matrices derived from human listeners. As with human listeners, nearly all of the confusions involve vowels that are quite close to one another in phonetic space; e.g., /ɑ/ is confused mainly with /ɔ/ and /ʌ/, /ɛ/ is confused mainly with /ɪ/ and /æ/, etc. There is also clear evidence that individual stimuli that were less intelligible to human listeners were far more likely to be misclassified by the narrow band recognizer. The error rate for the narrow band model for the 192 tokens with listener error rates of 15% or higher

TABLE I. City-block spectral distances between a sequence of spectra for a sample input signal (/hæd/, spoken by a man) and templates for each of 12 vowels. The first five rows show spectral distances sampled at equally spaced points from 15% to 75% to of vowel duration; the last row shows the weighted mean of these five distances, using weights of 1.0, 1.0, 1.0, 1.0, and 0.6 for the distances computed at 15%, 30%, 45%, 60%, and 75% of vowel duration, respectively. The minimum distance in each row is shown in bold.

	Vowel											
Time slice	/i/	/ɪ/	/e/	/ɛ/	/æ/	/ɑ/	/ɔ/	/o/	/ʊ/	/u/	/ʌ/	/ɝ/
Slice 1 (15%)	1.071	0.970	0.962	**0.871**	0.872	0.963	0.950	0.950	0.973	1.028	0.929	0.946
Slice 2 (30%)	1.079	0.964	1.012	**0.858**	0.872	0.953	0.948	0.968	0.986	1.034	0.926	0.973
Slice 3 (45%)	1.068	0.945	1.032	**0.879**	0.881	0.963	0.963	0.990	0.989	1.028	0.945	0.978
Slice 4 (60%)	1.091	0.950	1.063	0.885	**0.868**	0.964	0.960	1.003	1.015	1.056	0.958	0.998
Slice 5 (75%)	1.094	0.986	1.079	0.930	**0.881**	0.911	0.928	1.006	0.990	1.055	0.942	0.994
Weighted mean	1.079	0.961	1.025	0.880	**0.874**	0.954	0.952	0.982	0.991	1.039	0.940	0.977

TABLE II. Confusion matrix relating the vowel intended by the talker to the vowel recognized by the narrow band classification model. Results are summed across the three talker groups. Values on the main diagonal, indicating correctly recognized stimuli, are shown in boldface.

	Vowel as classified by the narrow band pattern recognition model											
	/i/	/ɪ/	/e/	/ɛ/	/æ/	/ɑ/	/ɔ/	/o/	/ʊ/	/u/	/ʌ/	/ɝ/
/i/	**93.5**	0.7	2.9						0.7	2.2		
/ɪ/		**93.5**	3.6	0.7					1.4	0.7		
/e/	3.5	2.1	**92.9**					0.7	0.7			
/ɛ/		4.3		**88.4**	4.4		0.7		0.7		1.4	
/æ/				13.8	**85.5**	0.7						
/ɑ/						**91.3**	6.5				2.2	
/ɔ/					0.7	11.4	**86.4**				1.5	
/o/			0.7				0.7	**94.3**	2.8	1.4		
/ʊ/			0.7					0.7	**94.2**	0.7	2.1	1.4
/u/	5.6	0.7				0.7		3.5	2.2	**87.2**		
/ʌ/						1.4	2.9		5.1		**90.6**	
/ɝ/		0.7		0.7					0.7			**97.8**

was 30.2%, more than five times greater than the error rate for the remaining well identified tokens (5.9%).[3]

Along with many similarities between the labeling patterns of humans and those of the narrow band model, there are some differences which reveal limitations of the current version of the model to simulate human vowel perception in detail. The most important difference is the occasional tendency of the model to confuse high front vowels with high back vowels. For example, note the ~3% of /i/ tokens that were incorrectly recognized as /u/ or /U/, the ~2% of /ɪ/ tokens that were incorrectly recognized as /u/ or /U/ and, most important, the ~6% of /u/ tokens that were incorrectly recognized as /i/ or /ɪ/. While these front–back confusions occurred on a small fraction of the utterances, this type of error is nearly nonexistent in human listener data (e.g., see Table VII in Hillenbrand *et al.*, 1995).

In keeping with a large body of evidence from human listeners, the narrow band model is better at recognizing vowels when spectral change is incorporated into the distance measure than when recognition is based on a single spectral slice. The five columns to the left in Table III show recognition rates by the model based on a single spectral slice sampled at either 15%, 30%, 45%, 60%, or 75% of vowel duration. These single-slice recognition rates, which typically vary from 75% to 80% correct, are considerably lower than the ~91% correct performance of the model that uses a sequence of five spectral slices.

The last three groups of columns in Table III compare versions of the recognition model incorporating two, three, and five spectral slices. The two-slice model used slices 1 and 5 with weights of 1 and 0.6, respectively, while the three-slice model used slices 1, 3, and 5 (15%, 45% and 75%) with weights of 1.0, 1.0, and 0.6. As the table shows, all of the multi-slice models performed better than the single-slice models, and performance differences among the two-, three-, and five-slice models are marginal at best. For the two-slice model, extensive testing with different combinations of slice locations showed a consistent advantage for more widely spaced slices; e.g., locations such as 1-5, 1-4, 2-5, etc., produced better recognition accuracy than locations such as 2-3, 2-4, etc. Variation in weighting scheme produced only slight differences in recognition accuracy. Similarly, the three-slice model performed better with more widely spread slice locations (e.g., 1-3-5, 1-3-4, 2-3-5, etc.) than tightly spaced locations (e.g., 1-2-3, 2-3-4, etc.), with little effect of weighting scheme. Taken as a whole, the results in Table III are consistent with data from both human listening studies and studies using pattern recognition methods indicating that spectral change patterns play a secondary but quite important role in vowel identification (for reviews, see Strange, 1989; Nearey, 1989; Hillenbrand and Nearey, 1999).

Table IV was designed to provide some insight into the benefit that is derived from the two major signal processing steps that are used to generate the test spectra and templates. The table shows recognition accuracy by the narrow band pattern matching model with and without thresholding and/or SLN. Results are for the five-slice version of the model de-

TABLE III. The five columns to the left of the table show the percent correct recognition accuracy by the narrow band model based on a single spectral slice, taken at 15% (slice 1), 30% (slice 2), 45% (slice 3), 60% (slice 4), or 75% (slice 5) of vowel duration. The three columns to the right of the table show percent correct figures for two-, three- and five-slice versions of the recognition model. Results are shown separately for syllables spoken by men, women, and children, with the bottom row showing a mean computed across the three talker groups.

	Slice no.					No. of slices		
	1	2	3	4	5	2	3	5
Men	79.8	83.7	85.4	82.2	74.4	94.3	93.1	92.0
Women	74.5	79.7	78.5	76.6	66.3	92.0	92.4	92.2
Children	72.1	75.2	77.2	72.6	65.9	85.5	89.3	90.0
Mean	75.5	79.5	80.4	77.1	68.9	90.6	91.6	91.4

TABLE IV. Recognition accuracy by the narrow band pattern-recognition model with and without thresholding and spectrum-level normalization (SLN).

	Thresholding and SLN	Thresholding, no SLN	No thresholding, SLN	No thresholding, No SLN
Men	92.0	92.0	60.0	50.9
Women	92.2	90.1	58.7	60.8
Children	90.0	86.4	61.1	57.8
Mean	91.4	89.5	59.9	56.5

scribed above. It can be seen that the broadband SLN operation, which was intended to reduce the importance of variation in formant amplitude relationships, provides at best a very small benefit. We retain the SLN operation because it does no harm, and it is also quite possible that the SLN operation would prove its usefulness in recognizing less well behaved test signals, such as those contrived by Klatt (1982a) in which spectral-shape features such as formant amplitude relations and spectral tilt have been deliberately altered from their typical values.

The most striking finding in Table IV is that the thresholding operation, implemented by zeroing out all spectral values below a 328-Hz Gaussian running average of spectral amplitudes, has a dramatic effect on the performance of the recognizer. When the thresholding operation is removed, the performance of the recognizer drops by over 30 percentage points. We assume that this operation improves recognition accuracy by suppressing spectral components in between harmonic peaks and in the valleys between formants, emphasizing those aspects of the spectrum that are most closely associated with vowel identity.

V. DISCUSSION

Our primary conclusion from these findings is that it is possible in principle to recognize vowels using a pattern-matching scheme involving the direct comparison of unsmoothed harmonic spectra with a set of empirically derived vowel templates. The overall recognition accuracy for the model approached but did not quite equal that of human listeners, and it is clear that at least part of the performance advantage for listeners can be attributed to listeners' use of duration cues. A goal of future work with this model is to develop some method to incorporate duration information along with the spectral distances that form the basis of the current method. On first glance this would seem to be a straightforward problem, but our earlier work on the use of duration cues by human listeners (Hillenbrand *et al.*, 2000a) showed that listeners are quite smart and flexible in their use of vowel duration. For example, listeners made virtually no use of duration cues in distinguishing pairs such as /i/−/ɪ/, /e/−/ɛ/, and /u/−/U/, in spite of large and systematic differences in duration separating these vowel pairs. Modeling work suggested that listeners assign little or no weight to duration for these vowel pairs because the vowels can be separated reliably on the basis of spectral cues alone. On the other hand, listeners make considerable use of duration cues in distinguishing vowels such as /æ/-/ɛ/ and the /ɑ/-/ɔ/-/ʌ/ cluster because these vowels show a greater degree of overlap in their spectral properties. We have experimented with some simple schemes for incorporating duration measures and have met with only modest success. We suspect that a humanlike scheme will be needed that assigns considerable weight to duration for some vowels and little or none for others.

It is of some interest to note that the recognition accuracy for the model was at most very slightly reduced at higher fundamental frequencies. For example, the model was as accurate in recognizing tokens spoken by women (92.2%) as men (92.0%), in spite of the roughly three-quarters of an octave difference in average fundamental frequency between the vowels of men ($\bar{\chi}=131$ Hz) and women ($\bar{\chi}=220$ Hz) in this database. This finding is consistent with labeling results from human listeners, who actually recognized syllables spoken by the women slightly (but significantly) better than those of the men (Hillenbrand *et al.*, 1995). There was a roughly 2 percentage point drop in recognition accuracy by the model for tokens produced by the children, which is consistent with a similar-size drop of 1–2 percentage points shown by our listeners. It is not entirely clear, for either the listeners or the model, whether the slightly lower intelligibility of the children's tokens has anything to do with their higher average fundamental frequencies ($\bar{\chi}=237$ Hz). It seems likely that the children's vowels were simply produced more variably and with a bit less articulatory precision than those of the adults (see Kent, 1976, for a review). Literature on the apparently simple question of whether vowel intelligibility degrades with increasing F_0 is surprisingly mixed. As noted, our study of naturally spoken /hVd/ syllables found no simple relationship between F_0 and vowel intelligibility. The same was true in a later study in which listeners identified formant synthesized versions of a 300-utterance subset of the same /hVd/ database (Hillenbrand and Nearey, 1999). The syllables were generated with a formant synthesizer driven by the original F_0 contours and either the original formant contours or flattened formant contours. Labeling results showed no evidence of a simple drop in intelligibility with increasing F_0. On the other hand, in Hillenbrand and Gayvert (1993), 300-ms signals with static formant patterns were synthesized based on F_0 and formant measurements of each of the 1520 signals in the Peterson and Barney (1952) database. There was a small but highly reliable drop in intelligibility with increasing F_0, whether the signals were generated with monotone pitch (men: 74.4%, women: 72.2%, children: 70.0%) or with falling pitch (men: 76.9%, women: 73.8%, children: 72.1%). However, even here the relationship between F_0 and intelligibility was

hardly simple since the roughly three-quarters of an octave difference in F_0 between men and women was accompanied by about the same small drop in intelligibility as the roughly one-quarter octave difference in F_0 between women and children. Ryalls and Liberman (1982) found that vowels with formants appropriate for a male talker were more intelligible at a 135-Hz F_0 than at 100 Hz, and that vowels with formants appropriate for a female talker were more intelligible at 185 Hz than at 250 Hz. Similarly, Sundberg and Gauffin (1982) found a decrease in intelligibility for vowels synthesized at F_0's between 260 and 700 Hz. For a nice summary and discussion of this issue, see de Cheveigné and Kawahara (1999).

This larger issue aside, for the test signals used here it is clear—for both listeners and the model—that vowel identity is conveyed quite well at both low and moderately high F_0, in spite of the fact that the template shape is much more sparsely sampled at higher fundamental frequencies. For the simple city-block distance method used by our model, the similarity that is measured between an input spectrum and a template will be dominated by spectral differences that are clearly irrelevant to vowel identity; that is, most of these channel-by-channel differences do not correspond to the voice-source harmonics which effectively sample the idealized envelope shape. Further, this problem will clearly be exacerbated at higher fundamental frequencies. A guiding assumption of the model is that these irrelevant differences will represent a more-or-less constant source of noise when comparing the input spectrum to all templates, meaning that the *variation* in the distance measure from one template to the next will be controlled mainly by spectral distances at the harmonics—this despite the fact that, unlike the *missing data model*, harmonics receive no special treatment. The relatively high classification accuracy by the model across a ~1-octave range of average fundamental frequencies and ~2-octave range of fundamental frequencies across individual tokens suggests that this assumption is valid, at least for this range of fundamental frequencies. Further, the fact that harmonics are not isolated for special treatment should in principle mean that the method ought to work without modification in the classification of whispered, breathy, or otherwise marginally periodic vowels. This remains to be tested.

While the narrow band model makes no use of fundamental frequency when computing token-to-template distance, pitch does, in fact, figure into the recognition process indirectly as a result of the strategy of using separate templates for signals spoken by men, women, and children. In our view, this method is defensible since the approach is consistent with a large body of psychological evidence suggesting that listener judgments of vowel timbre are, in fact, affected in an orderly way by F_0. For example, a large number of studies have demonstrated that when F_0 is increased, vowel quality can be maintained only by increasing formant frequencies (e.g., Potter and Steinberg, 1950; Miller, 1953; Fujisaki and Kawashima, 1968; Slawson, 1967; Carlson *et al.*, 1975; Ainsworth, 1975; Nearey, 1989). This finding implies that different standards are employed in evaluating the spectrum envelope at different fundamental frequencies. Further, formant-based modeling studies have shown that vowels can be classified with greater accuracy when F_0 is included among the classification features (e.g., Assmann *et al.*, 1982; Hirahara and Kato, 1992; Hillenbrand and Gayvert, 1993). Finally, a recent study by Scott *et al.* (2001) presented listeners with synthetic utterances generated by a source-filter vocoder in which the spectrum envelopes, the fundamental frequencies, or both the spectrum envelopes and the fundamental frequencies were shifted upward in frequency by varying amounts. The authors reported that high intelligibility could be maintained for shifted envelopes, but only if F_0 was also shifted up in frequency. Conversely, upward shifts in F_0 degraded intelligibility unless the F_0 shifts were accompanied by upward shifts in the spectrum envelope. Scott *et al.* argued that their results were best explained by assuming that listener judgments of phonetic quality were influenced by learned associations between F_0 and the spectrum envelope. Exactly how this kind of F_0-dependency is realized in the human system is unclear, and it is unlikely that it is anything as mechanical as the three-template method that we adopted. However, we would argue that our approach is broadly compatible with well-established findings on the interaction between F_0 and the spectral envelope in the perception of vowel quality.

There are several aspects of the labeling behavior of the narrow band model which appear to match well-known characteristics of vowel classification by human listeners. For example, implementations of the model which incorporated spectral change classified vowels with substantially greater accuracy than single-slice implementations. This finding is consistent with a substantial body of evidence implicating a role for spectral change (e.g., Strange *et al.*, 1983; Nearey and Assmann, 1986; Jenkins *et al.*, 1983; Parker and Diehl, 1984; Andruski and Nearey, 1992; Jenkins and Strange, 1999; Hillenbrand and Gayvert, 1993; Hillenbrand and Nearey, 1999; Assmann and Katz, 2000, 2001). The single-slice recognition rates produced by the model vary from about 70% to 80%, depending on the location of the slice. These figures are quite similar to recognition rates reported in several studies in which human listeners were asked to identify either synthetic or naturally spoken vowels with static spectral patterns. For example, Hillenbrand and Gayvert (1993) reported a 74.8% identification rate for 300-ms steady-state synthetic vowels that were synthesized using the F_0 and formant measurements from the 1520-stimulus vowel database recorded by Peterson and Barney (1952). Similarly, Hillenbrand and Nearey reported a 73.8% identification rate for flat-formant resynthesized versions of 300 /hVd/ utterances drawn from the Hillenbrand *et al.* (1995) database (see also Assmann and Katz, 2000, 2001). Finally, Fairbanks and Grubb (1961) reported a 72.1% identification rate for naturally spoken static vowels, demonstrating that the relatively low intelligibility of static vowels is not an artifact of the synthesis methods used in Hillenbrand–Gayvert, Hillenbrand–Nearey, and Assmann–Katz studies.

Implementations of the model incorporating multiple slices of the spectrum produced substantially greater recognition accuracy than the static model, with average recognition rates of 90.6%, 91.6%, and 91.4% for two-, three-, and five-slice models, respectively. It is of some interest to note

that two samples of the spectrum work nearly as well as three or five, suggesting that it is the gross trajectory of spectral movement that is critical and not the fine details of spectral change. A similar conclusion was reached in our earlier modeling study (Hillenbrand *et al.*, 1995) using quadratic discriminant analysis for classification and formants (with or without F_0) as features. The results showed nearly identical classification rates for a discriminant classifier trained and tested on three samples of the formant pattern (20%–50%–80% of vowel duration) as compared to a two-sample model (20%–80% of vowel duration). Both sets of findings are consistent with the dual target model of vowel recognition proposed by Nearey and Assmann (1986). The present results, of course, are based on vowels in a fixed /hVd/ environment. It remains to be determined whether a simple two-sample specification of spectral movement is adequate for more complex utterances involving variation in the phonetic environment surrounding the vowel.

Beyond the straightforward calculation of a high resolution Fourier spectrum, the key signal processing operations in the narrow band model are quite simple, consisting of just two steps: (1) a SLN operation designed to flatten the spectrum and reduce the importance of formant amplitude differences, and (2) a thresholding operation designed to emphasize information-bearing spectral peaks at the expense of perceptually less relevant spectral regions in the narrow valleys between harmonics, and in the broad valleys between formant peaks. One of the more striking aspects of the findings was the dramatic effect that the thresholding operation had on the performance of the recognizer, with recognition accuracy falling from near-human levels of ~90%–92% to about 60% with the removal of the thresholding operation. Recall that the simple city-block method that is used to measure token-template distance treats the information-bearing harmonics in the same way as the perceptually irrelevant nonharmonic spectral values. In all cases, the sum of the token-template differences for nonharmonic spectral values will greatly exceed the sum of the distances for the perceptually relevant spectral values at harmonic frequencies for the simple reason that the great majority of the 256 spectral values do not correspond to harmonics. As noted above, the simplified similarity measure relies on the assumption that the irrelevant spectral distances will represent a more-or-less constant source of noise across all templates, thereby allowing differences at harmonic frequencies to account for most of the variation in the distance measure across templates. The dramatic improvement in recognition accuracy attributable to the thresholding operation indicates that this assumption is valid only if steps are taken to emphasize harmonics at the expense of nonharmonic values, and formants over the valleys between formants. A similar, independently developed peak-enhancement method developed by Liénard and Di Benedetto (2000) has been used with some success to recognize French vowels from smoothed (as opposed to narrow band) ''bump vectors''—smooth spectra resulting from a thresholding operation similar to the one described here. In keeping with the present findings, recognition experiments showed a substantial advantage for the bump vector over a variety of alternative smoothed spectral representations that did not incorporate a thresholding operation (see also Aikawa *et al.*, 1993).

In sharp contrast to the near-essential status of the thresholding operation, the influence of the SLN operation on recognition performance was marginal at best. It is not entirely clear what should be concluded from this. There are several reasons why the SLN operation may have had little or no impact on the performance of the recognition model. First, as we indicated above, it is possible that the test signals, which were all recorded at comfortable vocal efforts and with the same recording equipment, were too well behaved to allow the SLN to have much effect. It is possible that the SLN operation would have had a more substantial effect if factors such as vocal effort (e.g., Liénard and Di Benedetto, 2000) or channel characteristics had varied across stimuli, or if steps had been taken to deliberately alter natural spectral-tilt or formant-level characteristics, as in Klatt's (1982a) study. However, there is some reason to believe that part of the problem is that we have simply not found the right way to implement the SLN operation. Recall that, along with the many similarities in vowel classification between listeners and the narrow band model, there was one aspect of model classification that did not resemble human listener behavior. The model showed some tendency to confuse high-front vowels such as /i/ and /ɪ/ with high back vowels such as /u/ and /U/. While this type of error was not common in the model output, it is almost never observed in human listeners. Examination of individual signals showed that these confusions tended to occur in cases in which the formant pattern of the input spectrum matched the correct template reasonably well, but differences in formant amplitude resulted in the incorrect template producing the smallest token-template distance. For example, there were tokens of /u/ with weak second formants, resulting in good matches to the incorrect /i/ template. This is exactly the situation for which the SLN operation was designed, and in examining individual cases we found that the effect of the SLN was in the right direction, but the change in amplitude relations was not large enough. We have found that simple modifications to the SLN function that are designed to produce a greater degree of flattening across the spectrum (e.g., using the inverse of the *squared* running average as the gain function) can virtually eliminate front–back confusions, but in rather limited experimentation we have not yet found a method that is free of undesirable side effects (e.g., increasing the number of confusions among back vowels). An appropriate solution to this problem may well require transformation to a nonlinear frequency scale, such as one based on the critical band or relationships between characteristic frequency and distance along the basilar membrane.[4]

As we indicated above, we regard the present findings as an existence proof, demonstrating that vowels can be recognized with a high degree of accuracy based on a distance metric which directly measures the similarity between an unsmoothed harmonic spectrum and a set of empirically derived smoothed vowel templates. The present work is, of course, a first step and the findings are necessarily preliminary. To cite just a few examples, it is presently unclear whether the method will provide satisfying results (1) with

stimuli representing a variety of syllable types as opposed to the constant /hVd/ stimuli used here (but see Hillenbrand *et al.*, 2000b, for promising results using formant-based modeling and a wide variety of consonant environments), (2) with whispered or breathy signals, (3) with stimuli such as those used by Klatt (1982a), which are designed to provide a systematic test of sensitivity to spectral shape details such as overall spectral tilt, spectral notches, formant bandwidth, and formant amplitude relationships, and (4) a range of fundamental frequencies exceeding the roughly 2-octave range represented by individual tokens in the present database. These and related questions are among the goals of future work with this method.

ACKNOWLEDGMENTS

This work was supported by a grant from the National Institutes of Health Grant No. R01-DC01661 to Western Michigan University. We are grateful to Michael Clark for comments on earlier drafts.

[1]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; i.e., 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 1 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. The smoothing window sizes here and elsewhere refer to the width of the individual rectangular windows used in each of the three averaging passes.

[2]The thresholding operation that is being used here is closely related to simultaneous masking, that is, the tendency of highly active neurons in one frequency region 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* in relation to region *b*, and (2) the distance (i.e., difference in frequency) between the two regions. Both of these features are captured by a center-weighted average (implemented here with a Gaussian-weighted running average); i.e., the masking function clearly reflects average amplitude, but the masking influence exerted by a particular frequency region falls off systematically with increasing distance. In light of the *phonetically based* modeling goals of this work, no attempt was made to accurately simulate physiologically or psychophysically observed masking phenomena. As noted in Sec. II A, a physiologically accurate simulation of masking would of necessity retain spectral features that contribute to the detailed timbre percept of a stimulus but which may have little or no bearing on vowel identity. Consequently, the thresholding process used here was designed with the goal of maximizing the spectral similarity of vowels that are labeled identically by listeners. It is, in fact, for this reason that we have adopted the somewhat awkward term *thresholding* instead of the term *masking* that we used in some of our other writings describing this type of process (e.g., Hillenbrand and Houde, 2002).

[3]It might be argued that the lower recognition rate for the more poorly identified tokens is due to the simple and uninteresting fact that these tokens were not used to create the templates. We tested this possibility by repeating the tests described above, but with templates created from all 1668 tokens. As with the earlier tests, the model showed a much higher error rate for the 192 signals that were more poorly identified by human listeners (25.0%) than the signals that were well identified by listeners (6.6%).

[4]Limited experimentation with a theoretically reasonable basilar membrane distance scale based on Greenwood (1990) produced disappointing recognition results. Given the limited time we have devoted to this effort to date, we do not conclude much of anything from these results. However, we think it is quite possible that determining how much weight should be assigned to spectral differences in different frequency regions will involve more than simply transforming to a tonotopic scale. For example, the relative importance that is assigned to the F_1 and F_2 regions is likely to depend on the usefulness of these two regions in distinguishing one vowel from another rather than simply the relative amounts of space that are occupied along the basilar membrane.

Aikawa, K., Singer, H., Kawahara, H., and Tohkura, Y. (**1993**). ''A dynamic cepstrum incorporating time-frequency masking and its application to continuous speech recognition,'' ICASSP-93, 668–671.

Ainsworth, S. (**1975**). ''Intrinsic and extrinsic factors in vowel judgments,'' in Auditory Analysis and the Perception of Speech, edited by G. Fant and M. A. A. Tathum (Academic, London).

Andruski, J., and Nearey, T. M. (**1992**). ''On the sufficiency of compound target specification of isolated vowels and vowels in /bVb/ syllables,'' J. Acoust. Soc. Am. **91**, 390–410.

Assmann, P., and Summerfield, Q. (**1989**). ''Modeling the perception of concurrent vowels: Vowels with the same fundamental frequency,'' J. Acoust. Soc. Am. **85**, 327–338.

Assmann, P., and Katz, W. (**2000**). ''Time-varying spectral change in the vowels of children and adults,'' J. Acoust. Soc. Am. **108**, 1856–1866.

Assmann, P., and Katz, W. (**2001**). ''Effects of synthesis fidelity on vowel identification: Role of spectral change and voicing source,'' J. Acoust. Soc. Am. **110**, 2658(A).

Assmann, P., Nearey, T., and Hogan, J. (**1982**). ''Vowel identification: orthographic, perceptual, and acoustic aspects,'' J. Acoust. Soc. Am. **71**, 975–989.

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, 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., Fant, G., and Granstrom, B. G. (**1975**). ''Two-formant models, pitch, and vowel perception,'' in *Auditory Analysis and Perception of Speech*, edited by G. Fant and M. A. A. Tathum (Academic, London), pp. 55–82.

Chistovich, L. A., and Lublinskaya, V. V. (**1979**). ''The 'center of gravity' effect in vowel spectra and critical distance between formants: Psychoacoustical study of the perception of vowel-like stimuli,'' Hear. Res. **1**, 185–195.

de Cheveigné, A., and Kawahara, H. (**1999**). ''A missing data model of vowel identification,'' J. Acoust. Soc. Am. **105**, 3497–3508.

Disner, S. F. (**1980**). ''Evaluation of vowel normalization procedures,'' J. Acoust. Soc. Am. **76**, 253–261.

Fairbanks, G., and Grubb, P. (**1961**). ''A psychophysical investigation of vowel formants,'' J. Speech Hear. Res. **4**, 203–219.

Fujisaki, H., and Kawashima, T. (**1968**). ''The roles of pitch and higher formants in the perception of vowels,'' IEEE Trans. Audio Electroacoust. **AU-16**, 73–77.

Greenwood, D. (**1990**). ''A cochlear frequency-position function for several species—29 years later,'' J. Acoust. Soc. Am. **87**, 2592–2605.

Hillenbrand, J., and Gayvert, R. T. (**1993**). ''Vowel classification based on fundamental frequency and formant frequencies,'' J. Speech Hear. Res. **36**, 694–700.

Hillenbrand, J. M., and Nearey, T. N. (**1999**). ''Identification of resynthesized /hVd/ syllables: Effects of formant contour,'' J. Acoust. Soc. Am. **105**, 3509–3523.

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

Hillenbrand, J. M., Clark, M. J., and Houde, R. A. (**2000a**). ''Some effects of duration on vowel recognition,'' J. Acoust. Soc. Am. **108**, 3013–3022.

Hillenbrand, J. M., Clark, M. J., and Nearey, T. M. (**2000b**). ''Effects of consonant environment on vowel formant patterns,'' J. Acoust. Soc. Am. **109**, 748–763.

Hillenbrand, J., Getty, L. A., Clark, M. J., and Wheeler, K. (**1995**). ''Acoustic characteristics of American English vowels,'' J. Acoust. Soc. Am. **97**, 3099–3111.

Hirahara, T., and Kato, H. (**1992**). ''The effect of *F*0 on vowel identification,'' in *Speech Perception, Production and Linguistic Structure*, edited by Y. Tohkura, E. Vatikiotis-Bateson, and Y. Sagisaka (Ohmsha, Tokyo), pp. 89–112.

Ito, M., Tsuchida, J., and Yano, M. (**2001**). ''On the effectiveness of whole spectral shape for vowel perception,'' J. Acoust. Soc. Am. **110**, 1141–1149.

Jenkins, J. J., and Strange, W. (**1999**). ''Perception of dynamic information for vowels in syllable onsets and offsets,'' Percept. Psychophys. **61**, 1200–1210.

Jenkins, J. J., Strange, W., and Edman, T. R. (**1983**). ''Identification of vowels in 'vowelless' syllables,'' Percept. Psychophys. **34**, 441–450.

Kent, R. D. (**1976**). ''Anatomical and neuromuscular maturation of the speech mechanism: Evidence from acoustic studies,'' J. Speech Hear. Res. **19**, 421–447.

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

Klatt, D. H. (**1982b**). ''Speech processing strategies based on auditory models,'' in *The Representation of Speech in the Peripheral Auditory System*, edited by R. Carlson and B. Granstrom (Elsevier Biomedical, Amsterdam), pp. 181–196.

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

Miller, G. A. (**1956**). ''The perception of speech,'' in *For Roman Jakobsen: Essays on the Occasion of his Sixtieth Birthday*, edited by M. Halle ('s-Gravenhage, Mouton, The Netherlands), pp. 353–359.

Miller, J. D. (**1989**). ''Auditory-perceptual interpretation of the vowel,'' J. Acoust. Soc. Am. **85**, 2114–2134.

Miller, R. L. (**1953**). ''Auditory tests with synthetic vowels,'' J. Acoust. Soc. Am. **18**, 114–121.

Nearey, T. M. (**1989**). ''Static, dynamic, and relational properties in vowel perception,'' J. Acoust. Soc. Am. **85**, 2088–2113.

Nearey, T. M. (**1992**). ''Applications of generalized linear modeling to vowel data,'' in *Proceedings of ICSLP 92*, edited by J. Ohala, T. Nearey, B. Derwing, M. Hodge, and G. Wiebe (University of Alberta, Edmonton, AB), pp. 583–586.

Nearey, T. M., and Assmann, P. (**1986**). ''Modeling the role of vowel inherent spectral change in vowel identification,'' J. Acoust. Soc. Am. **80**, 1297–1308.

Nearey, T. M., Hogan, J., and Rozsypal, A. (**1979**). ''Speech signals, cues and features,'' in *Perspectives in Experimental Linguistics*, edited by G. Prideaux (Benjamin, Amsterdam), pp. 73–96.

Parker, E. M., and Diehl, R. L. (**1984**). ''Identifying vowels in CVC syllables: Effects of inserting silence and noise,'' Percept. Psychophys. **36**, 369–380.

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

Pols, L. C. W., van der Kamp, L. J., and Plomp, R. (**1969**). ''Perceptual and physical space of vowel sounds,'' J. Acoust. Soc. Am. **46**, 458–467.

Potter, R. K., and Steinberg, J. C. (**1950**). ''Toward the specification of speech,'' J. Acoust. Soc. Am. **22**, 807–820.

Remez, R. E., Rubin, P. E., Pisoni, D. B., and Carrell, T. E. (**1981**). ''Speech perception without traditional speech cues,'' Science **212**, 947–950.

Rosner, B. S., and Pickering, J. B. (**1994**). *Vowel Perception and Production* (Oxford U.P., Oxford).

Ryalls, J. H., and Liberman, A. M. (**1982**). ''Fundamental frequency and vowel perception,'' J. Acoust. Soc. Am. **72**, 1631–1634.

Scott, J. M., Assmann, P. F., and Nearey, T. N. (**2001**). ''Intelligibility of frequency-shifted speech,'' J. Acoust. Soc. Am. **109**, 2316(A).

Slawson, A. W. (**1967**). ''Vowel quality and musical timbre as functions of spectrum envelope and fundamental frequency,'' J. Acoust. Soc. Am. **43**, 87–101.

Strange, W. (**1989**). ''Dynamic specification of coarticulated vowels spoken in sentence context,'' J. Acoust. Soc. Am. **85**, 2135–2153.

Strange, W., Jenkins, J. J., and Johnson, T. L. (**1983**). ''Dynamic specification of coarticulated vowels,'' J. Acoust. Soc. Am. **74**, 695–705.

Sundberg, J., and Gauffin, J. (**1982**). ''Amplitude of the fundamental and the intelligibility of super pitch sung vowels,'' in *The Representation of Speech in the Peripheral Auditory System*, edited by R. Carlson and B. Granstrom (Elsevier Biomedical, Amsterdam), pp. 223–238.

Syrdal, A. K., and Gopal, H. S. (**1986**). ''A perceptual model of vowel recognition based on the auditory representation of American English vowels,'' J. Acoust. Soc. Am. **79**, 1086–1100.

Zahorian, S., and Jagharghi, A. (**1986**). ''Matching of 'physical' and 'perceptual' spaces for vowels,'' J. Acoust. Soc. Am. Suppl. 1 **79**, S8.

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