Identification of resynthesized /hVd/ utterances: Effects of formant contour

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

Terrance M. Nearey
Department of Linguistics, University of Alberta, Edmonton, Alberta T6G 2E7, Canada

(Received 15 May 1998; revised 1 November 1998; accepted 10 March 1999)

The purpose of this study was to examine the role of formant frequency movements in vowel recognition. Measurements of vowel duration, fundamental frequency, and formant contours were taken from a database of acoustic measurements of 1668 /hVd/ utterances spoken by 45 men, 48 women, and 46 children [Hillenbrand *et al.*, J. Acoust. Soc. Am. **97**, 3099–3111 (1995)]. A 300-utterance subset was selected from this database, representing equal numbers of 12 vowels and approximately equal numbers of tokens produced by men, women, and children. Listeners were asked to identify the original, naturally produced signals and two formant-synthesized versions. One set of ''original formant'' (OF) synthetic signals was generated using the measured formant contours, and a second set of ''flat formant'' (FF) signals was synthesized with formant frequencies fixed at the values measured at the steadiest portion of the vowel. Results included: (a) the OF synthetic signals were identified with substantially greater accuracy than the FF signals; and (b) the naturally produced signals were identified with greater accuracy than the OF synthetic signals. Pattern recognition results showed that a simple approach to vowel specification based on duration, steady-state F_0, and formant frequency measurements at 20% and 80% of vowel duration accounts for much but by no means all of the variation in listeners' labeling of the three types of stimuli. © *1999 Acoustical Society of America.* [S0001-4966(99)04406-9]

PACS numbers: 43.72.Ar, 43.71.Es, 43.72.Ja [JH]

INTRODUCTION

There is a long tradition of representing vowels by a single spectral cross section taken from the nucleus of the vowel. The essence of this approach was summarized nicely by Tiffany (1953):

> It has been commonly assumed or implied that the essential physical specification of a vowel phoneme could be accomplished in terms of its acoustic spectrum as measured over a single fundamental period, or over a short interval including at most a few cycles of the fundamental frequency. That is to say, each vowel has been assumed to have a unique energy vs frequency distribution, with the significant physical variables all accounted for by an essentially cross-sectional analysis of the vowel's harmonic composition. (p. 290).

The potential limitations of this static approach to vowel quality were recognized by Tiffany, who noted that vowel duration and changes over time in formant frequencies and the fundamental frequency (F_0) may play a role in vowel perception (see also similar comments by Potter and Steinberg, 1950; Peterson and Barney, 1952; and Stevens and House, 1963). While the role of F_0 contour and duration received a fair amount of attention in early accounts of vowel recognition (e.g., Ainsworth, 1972; Bennett, 1968; Black, 1949; Stevens, 1959; Tiffany, 1953), it has only been more recently that the role of formant frequency movements has been examined systematically.

Evidence from several studies suggests that formant frequency movements do, in fact, play an important role in vowel perception. For example, Strange *et al.* (1983) and Nearey and Assmann (1986) showed high identification rates for ''silent center'' stimuli in which the vowel centers were gated out, leaving only brief onglides and offglides (see also Jenkins *et al.*, 1983; Parker and Diehl, 1984). Nearey and Assmann (1986) showed that it was not simply spectral movement that was required, but a specific pattern of spectral change throughout the course of the vowel. Brief excerpts of naturally produced vowels excised from nucleus and offglide segments were presented to listeners in three conditions: (1) natural order (nucleus followed by offglide); (2) repeated nucleus (nucleus followed by itself); and (3) reverse order (offglide followed by nucleus). Identification error rates for the natural-order signals were comparable to those for the original, unmodified vowels, while the repeated-nucleus and reverse-order conditions produced much higher error rates.

There is also evidence that vowels with static formant patterns are not particularly well identified. Hillenbrand and Gayvert (1993a) synthesized 300-ms monotone vowels with stationary formant patterns from the F_0 and formant measurements of each of the 1520 tokens in the Peterson and Barney (1952) /hVd/ database (2 repetitions of 10 vowels spoken by 33 men, 28 women, and 15 children). The 27% identification error rate for these steady-state synthetic signals was nearly five times greater than the error rate reported by Peterson and Barney for the original utterances. Synthesizing the signals with a falling pitch contour resulted in a highly significant but relatively small drop in the error rate. These results suggest that the duration and spectral change information that was removed from these stimuli by the steady-state synthesis method plays an important role in vowel recognition. Similarly high identification error rates were reported by Fairbanks and Grubb (1961) for naturally produced sustained vowels.

Evidence implicating a role for spectral change also comes from studies using statistically based pattern classifiers. For example, Assmann *et al.* (1982) trained a linear discriminant classifier with: (a) steady-state F_0 and formant information alone; and (b) steady-state information plus formant slopes and duration. The pattern classification model that incorporated dynamic information provided more accurate predictions of error patterns produced by human listeners (see also Nearey and Assmann, 1986; Parker and Diehl, 1984). Hillenbrand *et al.* (1995) trained a quadratic discriminant classifier on F_0 and formant measurements from /hVd/ utterances spoken by 45 men, 48 women, and 46 children. The pattern classifier was trained on various combinations of acoustic measurements, with formant frequencies sampled either at steady state or at 20% and 80% of vowel duration. The classifier was much more accurate when it was trained on two samples of the formant pattern. For example, with F_1 and F_2 alone, the classification accuracy was 71% for a single sample at steady state and 91% for two samples of the formant pattern. Two-sample parameter sets including either F_0 or F_3 produced classification accuracies approaching those of human listeners. Adding vowel duration to the parameter set also improved classification accuracy, although the effect was relatively small if the formant pattern was sampled twice.

While the pattern classification evidence is clearly relevant, it needs to be kept in mind that demonstrating that a pattern classifier is strongly affected by spectral change is not the same as showing that human listeners rely on spectral change patterns. As Nearey (1992) noted, pattern classification results have only an indirect bearing on perception unless the classification results are compared to human listener data (see also Shankweiler *et al.*, 1977). A particularly clear example of the limitations of pattern classification studies can be seen by comparing pattern classification results using static acoustic measurements with studies in which human listeners identify vowels with static formant patterns. Several pattern classification studies have shown that vowels can be identified with relatively modest error rates in the 12%–14% range based on static acoustic measurements (e.g., Hillenbrand and Gayvert, 1993b; Miller, 1984, 1989; Nearey, 1992; Nearey *et al.*, 1979; Syrdal and Gopal, 1986). However, Hillenbrand and Gayvert (1993a) reported a 27% error rate for human listeners who were asked to identify static vowels synthesized from the Peterson and Barney (1952) F_0 and formant measurements, and Fairbanks and Grubb (1961) reported a 26% error rate for naturally produced static vowels. The Fairbanks and Grubb findings are especially striking since there were just seven talkers, all of them men, and the investigators went to great lengths to ensure the quality and representativeness of their test signals.

In our view, the primary lesson from this mismatch of pattern recognition results and listening tests is not that pattern classification evidence is necessarily irrelevant or misleading, but rather, that pattern recognition studies need to be followed up with appropriately designed perception experiments. The purpose of the present experiment was to follow up on the pattern classification tests reported in Hillenbrand *et al.* (1995) by asking listeners to identify naturally produced /hVd/ signals and two synthetically generated versions. One set of synthesized signals was generated using the

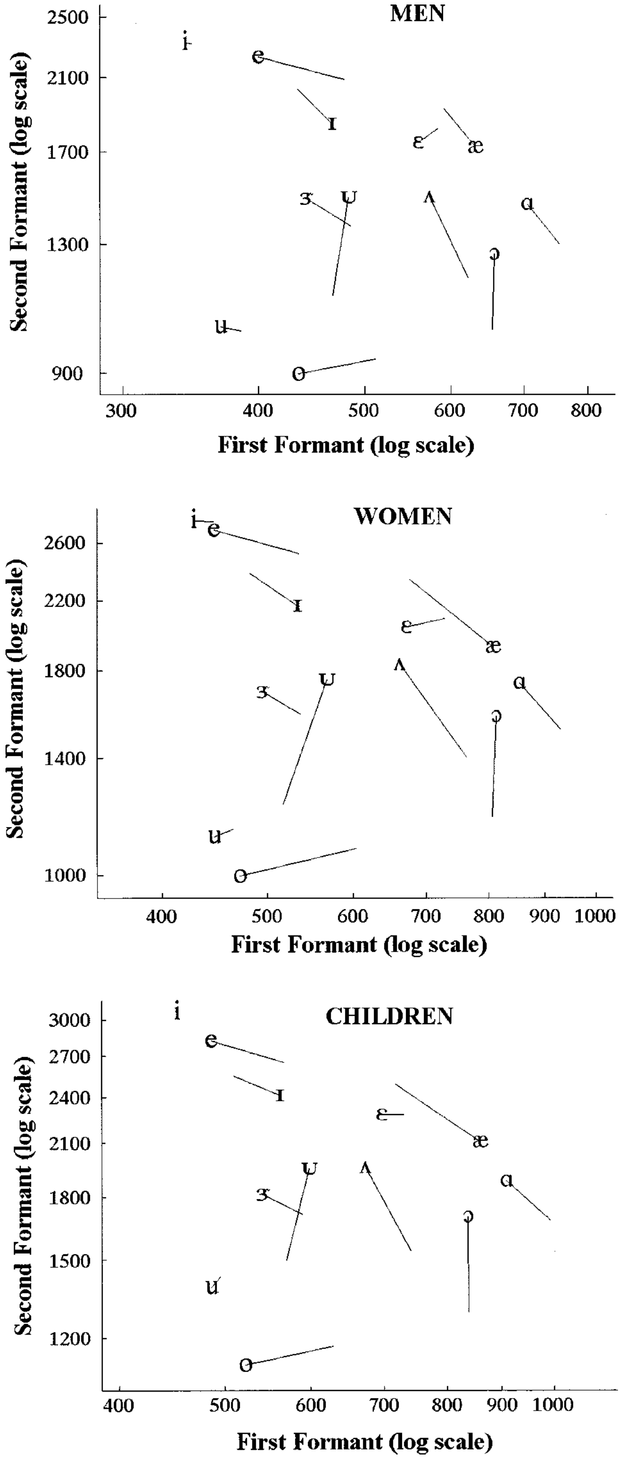

FIG. 1. Spectral change patterns for /hVd/ utterances produced by men, women, and children. The symbol identifying each vowel is plotted at the F_1–F_2 value for the second sample of the formant pattern (80% of vowel duration), and a line connects this point to the first sample (20% of vowel duration). The measurements are from Hillenbrand *et al.* (1995).

original measured formant contours and a second set of signals was synthesized with flat formants, fixed at the values measured at the steadiest portion of the vowel.

I. EXPERIMENT 1. METHODS

A. Test signals

The test signals consisted of a 300-stimulus subset of the 1668 /hVd/ utterances recorded by Hillenbrand *et al.* (1995). The talkers in that study consisted of 45 men, 48 women, and 46 10- to 12-year-old children. Recordings were made of subjects producing the vowels /i, ɪ, e, ɛ, æ, ɑ, ɔ, o, ʊ, u, ʌ, ɝ/ in /hVd/ syllables. A computer program was written to select a 300-stimulus subset from the full set of 1668 signals. The 300 signals were selected at random, but with the following constraints: (a) signals showing formant mergers involving any of the three lowest formants were omitted; (b) signals with identification error rates (measured in the original 1995 study) of 15% or greater were omitted; and (c) all 12 vowels were equally represented. The 300-stimulus set that was selected by this method included tokens from 123 of the 139 talkers, with 30% of the tokens from men, 36% from women, and 34% from children.

B. Acoustic measurements

Acoustic measurement techniques are described in detail in Hillenbrand *et al.* (1995). Briefly, peaks were extracted from LPC spectra every 8 ms and formant contours for F_1–F_4 were edited by hand during the vowel using a custom interactive editing tool. Measurements were also made of F_0 contour (also edited by hand using the same editing tool) and three temporal quantities: (a) the onset of the vowel; (b) the offset of the vowel; and (c) steady-state time; i.e., the single frame at which the formant pattern was judged by visual inspection to be maximally steady. Vowel onsets and offsets were also judged by visual inspection, using standard measurement criteria (Peterson and Lehiste, 1960). Average spectral change patterns for the full data set are shown in Fig. 1. The figure was created by connecting a line between the formant frequencies sampled at 20% and 80% of vowel duration; the symbol for each vowel category is plotted at the location of the second measurement. The measurement results are described in some detail in Hillenbrand *et al.*, but there are two points about the formant-change patterns in Fig. 1 that are particularly relevant to the present study. First, with the exception of /i/ and /u/, the formant frequency values show a good deal of change throughout the course of the vowel. For example, note that the vowels /e/ and /o/, which are known to be diphthongized, do not show spectral change magnitudes that are unusually large relative to the other vowels. Second, the formants change in such a way as to enhance the contrast between vowels with similar formant patterns. For example, the pairs /æ/–/ɛ/ and /u/–/ʊ/, which show a good deal of overlap when the vowels of individual talkers are plotted in static F_1–F_2 space (see Fig. 4 of Hillenbrand *et al.*), show very different patterns of formant-frequency change.

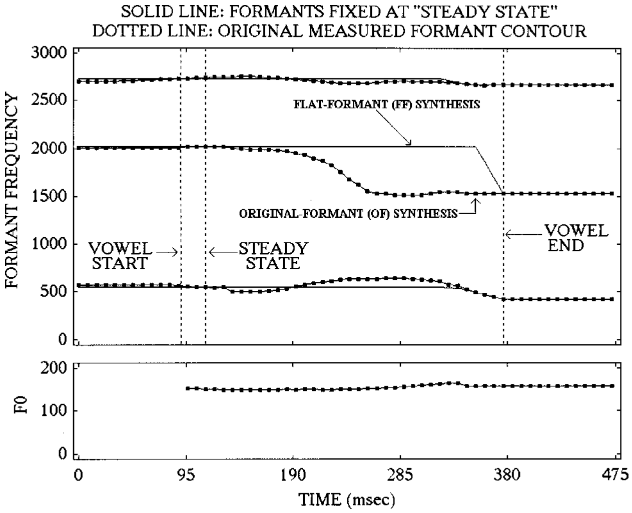

FIG. 2. Schematic spectrograms illustrating the method that was used to synthesize the original-formant (OF) and flat-formant (FF) signals.

C. Synthesis method

Test signals consisted of the 300 original, 16-kHz digitized utterances and two synthesized versions, for a total of 900 signals. The Klatt and Klatt (1990) formant synthesizer, running at a 16-kHz sample rate, was used to generate two sets of synthetic signals. The ''original formant'' (OF) and ''flat format'' (FF) synthesis methods are illustrated in Fig. 2. The OF signals were synthesized using the original measured formant contours, shown by the dashed curves. The vowel in this example is /æ/, and the measured formant contour shows a pronounced offglide which is quite common in our data for this vowel. The FF signals were synthesized with flat formants, fixed at the values measured at steady state, shown by the solid curves in Fig. 2.[1] For these signals a 40-ms linear transition connected the steady-state values to the F_1–F_3 values that were measured at the end of the vowel. Both sets of synthetic signals were generated with the measured F_0 contours and at their measured durations. During the /h/ interval (i.e., between the beginning of the stimulus and the start of the vowel): (a) the voicing amplitude (AV) parameter was set to zero and the aspiration amplitude (AH) parameter was controlled by the measured rms intensity of the signal being synthesized; (b) the F_1 bandwidth was set to 300 Hz; and (c) formant values for F_1–F_3 were set to the values that were measured at the start of the vowel. At all other times formant bandwidths remained at their default values ($B_1=90$, $B_2=110$, $B_3=170$, $B_4=400$, $B_5=500$, $B_6=800$). During the vowel the AH parameter was set to zero and the AV parameter was driven by the measured rms energy of the signal. Values of F_4 were set separately for each vowel and talker group based on data from Hillenbrand *et al.* (1995). Values of F_5 and F_6 were set separately for each talker group based on data from Rabiner (1968). Formant amplitudes were set automatically during the /h/ and vowel by running the synthesizer in cascade mode. A final /d/ was simulated by: (a) ramping F_1 100 Hz below its measured value at the end of the vowel in four steps of 25 Hz; and (b) switching from the cascade to the parallel branch of the synthesizer and setting the resonator

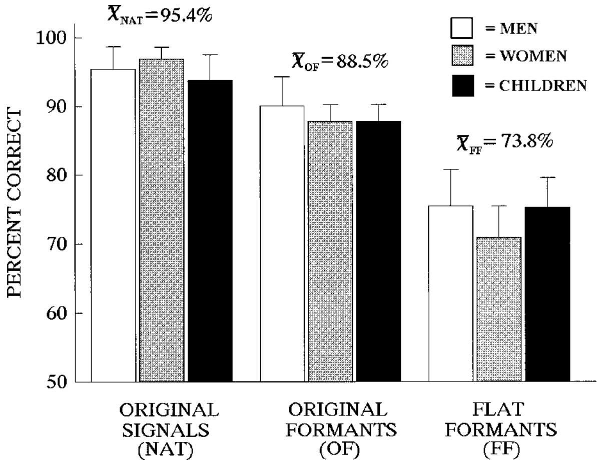

FIG. 3. Percent correct identification of the naturally produced signals, the original-formant synthetic signals, and the flat-formant synthetic signals. Error bars indicate one standard deviation. The percentages at the top indicate the mean percent correct for each condition pooled across the three talker groups.

gains of F_2–F_6 30 dB below the F_1 resonator gain; i.e., producing a ''voice bar'' with energy primarily at F_1. Since we were not entirely satisfied with our efforts to generate natural sounding final release bursts with the synthesizer, the signals were generated unreleased, and release bursts that had been excised from naturally produced signals spoken by one man, one woman, and one child were appended to the end of the stimuli.[2]

D. Listening Test

Twenty subjects who had taken or were currently enrolled in an undergraduate course in phonetics served as listeners. The choice of listeners with training in phonetic transcription was motivated by the findings of Assmann *et al.* (1982) indicating that many of the identification errors made by untrained subjects are due to listeners' uncertainty about how to map perceived vowel quality onto orthographic symbols. Subjects were tested one at a time in a quiet room in a single session lasting about 1 h. Listeners identified each of the 900 test signals (300 original signals, 300 OF signals, and 300 FF signals) presented in random order. The presentation order was reshuffled prior to each listening session. Stimuli were low-pass filtered at 6.9 kHz, amplified, and delivered at approximately 75 dBA over a single loudspeaker (Boston Acoustics A60) positioned approximately 1 m from the subject's head. Subjects entered their responses on a computer keyboard labeled with both phonetic symbols and key words for the 12 vowels. Subjects were allowed to repeat stimuli as many times as they wished before entering a response.

II. EXPERIMENT 1. RESULTS AND DISCUSSION

Figure 3 shows overall percent correct for each stimulus type and talker group averaged across all vowels. It can be seen that the naturally produced signals (NAT) were identified with the greatest accuracy, followed by the OF and FF synthesized signals. A two-way repeated-measures ANOVA using arcsine-transformed percent correct values showed a significant effect for stimulus type [$F(2,38)=860.7$, $p<0.0001$] and talker group [$F(2,38)=3.5$, $p<0.05$] and a significant interaction [$F(4,76)=10.2$, $p<0.0001$].[3] Newman–Keuls *post hoc* tests showed significant differences among all three stimulus types. Although statistically reliable, the effects for talker group are relatively small and nonuniform across stimulus type, as revealed by the significant interaction. For the naturally produced signals, *post hoc* analyses showed that the men's and women's tokens were identified with greater accuracy than the children's tokens. The pattern was different for the OF synthetic signals, which showed greater intelligibility for the men's tokens as compared to those of the women and children. A third pattern was observed for the FF synthetic signals, which showed significantly poorer intelligibility for the women's tokens as compared to the men and the children. One clear conclusion from this rather mixed set of talker-group results is that there was no evidence for a simple pitch effect; that is, although the spectrum envelope is more poorly defined at higher fundamental frequencies, the talker-group effects show no evidence of a simple inverse relationship between F_0 and vowel intelligibility [see also Carlson *et al.* (1975), and Hillenbrand and Gayvert (1993a), for related findings].

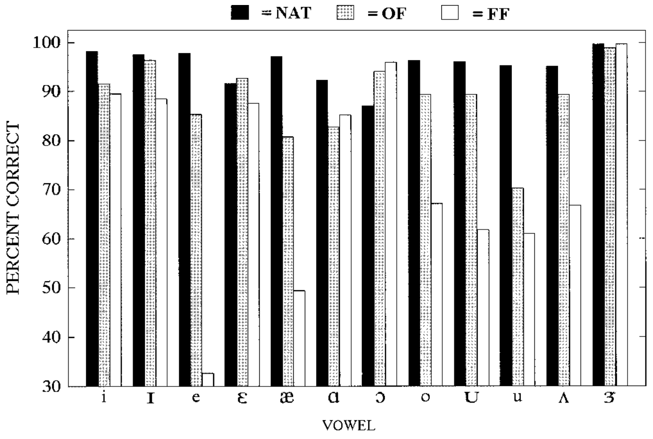

FIG. 4. Percent correct identification of the naturally produced signals (NAT), the original-formant (OF) synthetic signals, and the flat-formant synthetic signals separated by vowel category.

Figure 4 shows percent correct separately for each vowel category. Confusion matrices for the three conditions are shown in Tables I–III. It can be seen in Fig. 4 that the effect of stimulus type varies considerably from one vowel to the next. This was confirmed by a two-way repeated-measures ANOVA (using arcsine-transformed percent correct values) which tested the effects of stimulus type (NAT versus OF synthesis versus FF synthesis) and vowel. The ANOVA showed a significant effect for stimulus type [$F(2,38)=799.9$, $p<0.001$] and vowel [$F(11,209)=28.0$, $p<0.001$] as well as a significant interaction [$F(22,418)=46.7$, $p<0.001$]. The nature of the interaction will be discussed and further analyzed below.

Note that the pattern for /ɔ/ appears to differ markedly from the other vowels since this vowel showed the highest recognition accuracy in the FF condition and the lowest accuracy for the naturally produced signals. Examination of the confusion matrices in Tables I–III suggests that the overall

TABLE I. Confusion matrix for the naturally produced signals. Values on the main diagonal, indicating the percentage of trials in which the listeners' responses matched the vowel intended by the talker, are shown in boldface. The response means given in the last row indicate the percentage of trials in which a given vowel was used as a listener response. Each vowel, as classified by the speaker's intention, was presented on 8.3% of the trials.

		Vowel identified by listener											
		/i/	/ɪ/	/e/	/ɛ/	/æ/	/ɑ/	/ɔ/	/o/	/ʊ/	/u/	/ʌ/	/ɝ/
Vowel intended by talker	/i/	**98.2**	0.8	0.4	0.6	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯
	/ɪ/	⋯	**97.6**	0.6	1.4	⋯	⋯	⋯	⋯	0.4	⋯	⋯	⋯
	/e/	0.2	0.2	**97.8**	1.6	⋯	⋯	⋯	⋯	0.2	⋯	⋯	⋯
	/ɛ/	⋯	⋯	0.2	**91.6**	8.0	⋯	⋯	⋯	⋯	⋯	0.2	⋯
	/æ/	⋯	⋯	⋯	2.0	**97.2**	0.4	0.4	⋯	⋯	⋯	⋯	⋯
	/ɑ/	⋯	⋯	⋯	⋯	3.2	**92.2**	4.0	⋯	⋯	⋯	0.6	⋯
	/ɔ/	⋯	⋯	⋯	0.2	0.8	9.6	**87.2**	0.4	0.6	⋯	1.6	⋯
	/o/	⋯	⋯	⋯	⋯	⋯	⋯	⋯	**96.4**	1.4	0.6	1.6	⋯
	/ʊ/	⋯	⋯	⋯	⋯	⋯	⋯	⋯	0.2	**96.2**	0.4	3.2	⋯
	/u/	⋯	⋯	⋯	⋯	⋯	⋯	⋯	1.6	3.0	**95.4**	⋯	⋯
	/ʌ/	⋯	⋯	⋯	⋯	⋯	2.0	1.0	⋯	1.4	0.4	**95.2**	⋯
	/ɜ/	⋯	⋯	⋯	0.2	⋯	⋯	⋯	⋯	⋯	⋯	⋯	**99.8**
Response means:		8.2	8.2	8.3	8.1	9.1	8.7	7.7	8.2	8.6	8.1	8.5	8.3

percent correct figures may be misleading in some respects. The last row in each of these tables gives the response means; that is, the percentage of trials in which a given vowel was used as a response. Since each of the 12 vowels was presented equally often, an ideal listener whose responses always agreed with the speaker's intention would use each symbol on 8.3% of the trials. Note that the percentage of /ɔ/ responses increases from 7.7% for the naturally produced signals, to 8.6% for the OF synthetic signals, to 11.2% for the FF synthetic signals. In other words, for reasons that are not clear, there is an increasing probability of a listener hearing /ɔ/ across these three conditions. Although the overall percent correct for /ɔ/ improves from natural speech to OF synthesis to FF synthesis, the probability of a correct response on trials in which /ɔ/ was used as a response declines from 94.1% for the natural signals to 91.1% for the OF synthetic signals to 83.7% for the FF synthetic signals.

In the analyses that follow, we will first consider the effect of flattening the formants, and then consider the differences in intelligibility between the natural signals and the OF synthetic signals.

A. Effects of formant flattening

The drop in intelligibility that occurs as a result of flattening the formants is in general quite large. However, as Fig. 4 shows, the effect varies considerably from one vowel to the next. This was confirmed by a two-way repeated-measures ANOVA comparing just the OF and FF conditions, which showed significant effects for stimulus type $[F(1,19) = 366.1, \; p < 0.001]$ and vowel $[F(11,209) = 48.4, p < 0.001]$, and a significant interaction $[F(11,209) = 44.0, p < 0.001]$. Vowels showing the largest changes in intelligibility as a result of formant flattening were /e/ (52.8%), /æ/ (31.4%), /ʊ/ (27.6%), /ʌ/ (22.6%), and /o/ (22.2%).

TABLE II. Confusion matrix for the original-formant (OF) synthetic signals. Values on the main diagonal, indicating the percentage of trials in which the listeners' responses matched the vowel intended by the talker, are shown in boldface. The response means given in the last row indicate the percentage of trials in which a given vowel was used as a listener response. Each vowel, as classified by the speaker's intention, was presented on 8.3% of the trials.

		Vowel identified by listener											
		/i/	/ɪ/	/e/	/ɛ/	/æ/	/ɑ/	/ɔ/	/o/	/ʊ/	/u/	/ʌ/	/ɝ/
Vowel intended by talker	/i/	**91.6**	4.6	3.6	0.2	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯
	/ɪ/	0.6	**96.4**	0.2	2.4	0.4	⋯	⋯	⋯	⋯	⋯	⋯	⋯
	/e/	7.6	4.8	**85.4**	2.0	⋯	0.2	⋯	⋯	⋯	⋯	⋯	⋯
	/ɛ/	⋯	⋯	⋯	**92.8**	5.6	⋯	⋯	⋯	⋯	0.2	1.4	⋯
	/æ/	0.4	⋯	0.2	18.6	**80.8**	⋯	⋯	⋯	⋯	⋯	⋯	⋯
	/ɑ/	⋯	⋯	⋯	0.6	6.4	**82.8**	6.6	⋯	⋯	⋯	3.0	0.6
	/ɔ/	⋯	⋯	⋯	⋯	⋯	4.2	**94.2**	0.4	⋯	⋯	1.2	⋯
	/o/	⋯	⋯	⋯	⋯	⋯	0.4	⋯	**89.4**	5.2	4.2	0.8	⋯
	/ʊ/	⋯	⋯	⋯	⋯	⋯	0.2	⋯	1.0	**89.4**	2.0	7.0	0.4
	/u/	⋯	⋯	⋯	⋯	⋯	0.2	⋯	11.4	17.6	**70.4**	0.4	⋯
	/ʌ/	⋯	⋯	⋯	⋯	0.6	2.0	2.6	0.8	4.6	⋯	**89.4**	⋯
	/ɜ/	⋯	⋯	⋯	1.0	⋯	⋯	⋯	⋯	⋯	⋯	⋯	**99.0**
Response means:		8.4	8.8	7.5	9.8	7.8	7.5	8.6	8.6	9.7	6.4	8.6	8.3

TABLE III. Confusion matrix for the flat-formant (FF) synthetic signals. Values on the main diagonal, indicating the percentage of trials in which the listeners' responses matched the vowel intended by the talker, are shown in boldface. The response means given in the last row indicate the percentage of trials in which a given vowel was used as a listener response. Each vowel, as classified by the speaker's intention, was presented on 8.3% of the trials.

		Vowel identified by listener											
		/i/	/ɪ/	/e/	/ɛ/	/æ/	/ɑ/	/ɔ/	/o/	/ʊ/	/u/	/ʌ/	/ɝ/
Vowel intended by talker	/i/	**89.6**	6.6	2.8	0.8	⋯	0.2	⋯	⋯	⋯	⋯	⋯	⋯
	/ɪ/	4.4	**88.6**	2.8	3.6	0.6	⋯	⋯	⋯	⋯	⋯	⋯	⋯
	/e/	12.0	29.4	**32.6**	21.6	4.4	⋯	⋯	⋯	⋯	⋯	⋯	⋯
	/ɛ/	0.2	0.2	0.8	**87.6**	9.6	⋯	⋯	⋯	1.0	⋯	0.4	0.2
	/æ/	⋯	1.2	7.4	42.0	**49.4**	⋯	⋯	⋯	⋯	⋯	⋯	⋯
	/ɑ/	⋯	⋯	⋯	0.4	5.2	**85.2**	7.4	0.8	0.4	⋯	0.4	0.2
	/ɔ/	⋯	⋯	⋯	⋯	⋯	2.2	**96.0**	1.8	⋯	⋯	⋯	⋯
	/o/	⋯	⋯	⋯	0.2	⋯	3.2	12.2	**67.2**	9.0	7.0	1.2	⋯
	/ʊ/	⋯	⋯	⋯	⋯	⋯	⋯	⋯	13.2	**61.8**	10.2	14.6	0.2
	/u/	⋯	⋯	⋯	0.2	⋯	⋯	0.2	21.2	16.4	**61.0**	0.8	0.2
	/ʌ/	⋯	⋯	⋯	⋯	0.4	11.2	18.6	2.2	0.8	⋯	**66.8**	⋯
	/ɜ/	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	0.2	⋯	**99.8**
Response means:		8.5	10.5	3.9	13.3	5.8	8.5	11.2	8.9	7.5	6.5	7.0	8.4

Table IV shows what kinds of changes in identification occurred. The analysis focused on instances in which a given listener identified the OF version of a signal correctly (i.e., as the vowel intended by the talker) but the FF version incorrectly. The symbols going down the rows indicate the vowel as classified both by the speaker's intention and the listener's labeling of the OF version; the columns show how the FF versions of these signals were identified. The last column shows the total number of changes in identification in which the OF version was heard as the intended vowel but the FF version was heard as some other vowel. The most frequently occurring changes in identification (shown in boldface in Table IV) involved /e/ shifting to /ɪ/ or /ɛ/, /æ/ shifting to /ɛ/, /ʊ/ shifting to /ʌ/, /o/, or /u/, /ʌ/ shifting to /ɑ/ or /ɔ/, and /o/ shifting to /ɔ/. Labeling changes involving these five vowels accounted for approximately three-quarters of all correct-to-incorrect vowel shifts. A similar analysis that included all shifts in vowel color between OF and FF signals, whether from correct to incorrect or otherwise, yielded a pattern of results that was quite similar to that shown in Table IV.

As might be expected, vowels that typically show relatively large amounts of spectral change tended to be more strongly affected by formant flattening. Figure 5 shows the relationship between the average magnitude of formant frequency change for each vowel and the total number of correct-to-incorrect changes in identification (i.e., the last column of Table IV). The magnitude of spectral change for each vowel category was represented as the average length of a vector connecting formant measurements sampled at 20% of vowel duration and 80% of vowel duration. The vector was drawn in a three-dimensional space consisting of log-transformed values of F_1, F_2, and F_3. As Fig. 5 shows, the vowels tend to cluster into one group in the lower left showing relatively little spectral change and few changes in labeling, and a second group in the upper right showing a good deal of spectral change and many shifts in labeling. The relationship is far from perfect, however. For example, the perceptual effect of formant flattening for /e/ is quite large, even though the magnitude of formant frequency change is relatively modest in relation to the other vowels. We experi-

TABLE IV. Changes in phonetic labeling for signals whose original-formant versions were identified correctly but whose flat-formant versions were identified as a vowel other than that intended by the talker. The most frequently occurring vowel shifts are shown in boldface.

		FF synthetic vowel identified as												
		/i/	/ɪ/	/e/	/ɛ/	/æ/	/ɑ/	/ɔ/	/o/	/ʊ/	/u/	/ʌ/	/ɝ/	Total
OF vowel identified as	/i/	⋯	23	9	2	⋯	1	⋯	⋯	⋯	⋯	⋯	⋯	35
	/ɪ/	20	⋯	10	11	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	41
	/e/	37	**116**	⋯	**105**	22	⋯	⋯	⋯	⋯	⋯	⋯	⋯	**280**
	/ɛ/	1	1	4	⋯	38	⋯	⋯	⋯	5	⋯	2	1	52
	/æ/	⋯	2	28	**157**	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	**187**
	/ɑ/	⋯	⋯	⋯	⋯	6	⋯	26	3	1	⋯	⋯	⋯	36
	/ɔ/	⋯	⋯	⋯	⋯	⋯	11	⋯	7	⋯	⋯	⋯	⋯	18
	/o/	⋯	⋯	⋯	1	⋯	16	**55**	⋯	27	26	5	⋯	**130**
	/ʊ/	⋯	⋯	⋯	⋯	⋯	⋯	⋯	**50**	⋯	**45**	**55**	1	**151**
	/u/	⋯	⋯	⋯	1	⋯	⋯	1	**57**	**41**	⋯	3	1	104
	/ʌ/	⋯	⋯	⋯	⋯	2	**51**	**79**	8	2	⋯	⋯	⋯	**142**
	/ɜ/	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	1	⋯	⋯	1

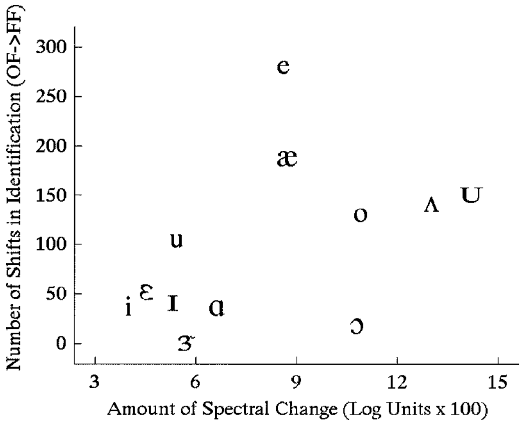

FIG. 5. Total number of correct-to-incorrect changes in identification as a function of the average magnitude of formant frequency change for each vowel. The magnitude of formant change for each vowel category was represented as the average length of a vector connecting formant measurements sampled at 20% of vowel duration and 80% of vowel duration. The vector was drawn in a three-dimensional space consisting of log-transformed values of F_1, F_2, and F_3.

mented with sample points other than 20% and 80%, and with a number of alternate methods of representing spectral change magnitude, including calculations based on normalization schemes proposed by Syrdal (1985); Syrdal and Gopal (1986); Miller (1984, 1989) and Peterson (1951). Results of these additional analyses were very similar to the general pattern shown in Fig. 5.

B. Natural signals versus OF synthesis

Although the main purpose of this study was to examine the effects of formant flattening, the difference in intelligibility that was observed between the naturally produced signals and the OF synthetic signals raises some important questions about the cues underlying the perception of vowel color. A two-way repeated-measures ANOVA comparing just the NAT and OF synthesis conditions showed significant effects for stimulus type $[F(1{,}19)=371.8,\ p<0.001]$ and vowel $[F(11{,}209)=10.6,\ p<0.001]$, and a significant interaction $[F(11{,}209)=21.0,\ p<0.001]$. Vowels showing the largest changes in intelligibility as a result of formant coding were /u/ (25.0%), /æ/ (16.4%), /e/ (12.4%), and /ɑ/ (9.4%).

Table V shows the distribution of responses for all instances in which a given listener identified the original signal correctly but identified the OF synthetic version as a vowel other than that intended by the talker. The most frequently occurring changes in identification involved /u/ shifting to /ʊ/ or /o/, /æ/ shifting to /ɛ/, and /e/ shifting to /i/. Labeling changes involving /u/ and /æ/ alone accounted for 37.8% of all correct-to-incorrect vowel shifts. These comparisons raise an obvious question: Why is there any difference in intelligibility between the natural signals and OF synthetic signals? In other words, what phonetically relevant information is not preserved by the formant frequency representation that drives the synthesizer during the vowel?

One possibility that cannot be ruled out is that the drop in intelligibility may occur at least in part as a result of errors in the estimation of formant frequencies. Measurement–remeasurement reliability for the LPC-derived formant frequency estimates is on the order of 1.0%–2.0% of formant frequency for F_1 and 1.0%–1.5% of formant frequency for F_2 and F_3 (Hillenbrand *et al.*, 1995).[4] However, these reliability estimates do not address the validity question, and the possibility exists that LPC produces systematic errors in formant frequency measurement. For example, in a relatively small-scale study, Di Benedetto (1989) reported LPC-derived estimates of F_2 and F_3 that were very similar to those derived from smoothed wide-band Fourier spectra, but estimates of F_1 that were systematically lower when measured with LPC. A larger and more formal comparison of 180 /hVd/ utterances by Hillenbrand *et al.* (1995) also found estimates of F_2 and F_3 that were similar between LPC and smoothed Fourier spectra; however, estimates of F_1 were found to be approximately 40 Hz *higher* for LPC.

The main question that is raised by these measurement issues is whether systematic errors in formant frequency estimation might account for the shifts in vowel quality that were observed between the natural and the OF synthetic signals. For example, the many labeling shifts that occurred from /u/ to /ʊ/ and /æ/ to /ɛ/ might be explained by positing that estimates of F_1 are systematically high. Figure 6 was

TABLE V. Changes in phonetic labeling for naturally produced signals that were identified correctly but whose original-formant synthetic versions were identified as a vowel other than that intended by the talker. The most frequently occurring vowel shifts are shown in boldface.

		OF synthetic vowel identified as												
		/i/	/ɪ/	/e/	/ɛ/	/æ/	/ɑ/	/ɔ/	/o/	/ʊ/	/u/	/ʌ/	/ɝ/	Total
NAT vowel identified as	/i/	⋯	21	16	1	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	38
	/ɪ/	3	⋯	⋯	12	2	⋯	⋯	⋯	⋯	⋯	⋯	⋯	17
	/e/	**38**	21	⋯	7	⋯	1	⋯	⋯	⋯	⋯	⋯	⋯	**67**
	/ɛ/	⋯	⋯	⋯	⋯	16	⋯	⋯	⋯	⋯	1	7	⋯	24
	/æ/	2	⋯	1	**89**	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	**92**
	/ɑ/	⋯	⋯	⋯	3	23	⋯	26	⋯	⋯	⋯	15	⋯	**67**
	/ɔ/	⋯	⋯	⋯	⋯	⋯	16	⋯	2	⋯	⋯	5	⋯	23
	/o/	⋯	⋯	⋯	⋯	⋯	2	⋯	⋯	22	17	2	⋯	43
	/ʊ/	⋯	⋯	⋯	⋯	⋯	1	⋯	5	⋯	10	27	⋯	45
	/u/	⋯	⋯	⋯	⋯	⋯	1	⋯	**54**	**80**	⋯	2	2	**137**
	/ʌ/	⋯	⋯	⋯	⋯	2	9	13	4	19	⋯	⋯	⋯	47
	/ɜ/	⋯	⋯	⋯	5	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	5

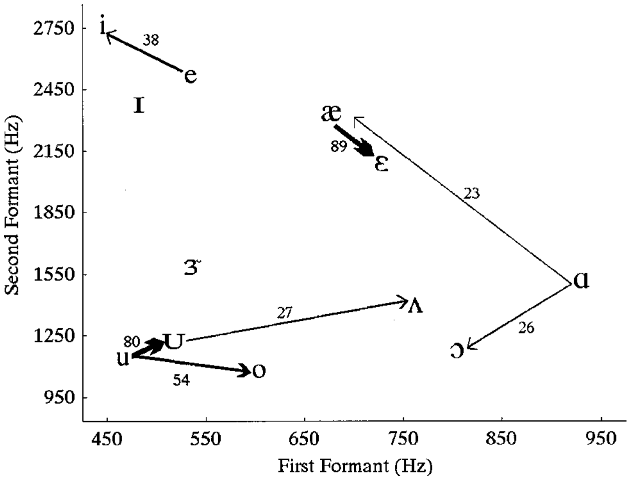

FIG. 6. This figure shows the most frequently occurring shifts in vowel color from NAT to OF; that is, the number of instances in which the naturally spoken version of an utterance was correctly identified but the original-formant synthetic version was identified as some other vowel. To provide a reference frame for formant space, phonetic symbols are plotted at the average values of F_1 and F_2 for the women. The tails of the arrows are plotted at the formant values for the correctly identified natural signal, and the arrow heads point at the phonetic label assigned to the OF synthetic version of the signal. The line thickness is roughly proportional to the number of labeling shifts.

designed to address this question. Plotted on this figure are the most frequently occurring shifts in vowel color from NAT to OF, based on the data in Table V. To provide a reference frame for formant space, phonetic symbols are plotted at the average steady-state values of F_1 and F_2 for the women. The tails of the arrows are plotted at the average formant values for the correctly identified natural signal, and the arrow heads point at the corresponding values for phonetic label assigned to the OF synthetic version of the signal. The line thickness (but *not* line length) is roughly proportional to the number of labeling shifts. The main point to be made about this figure is that no simple, systematic measurement error can account for the most common shifts in vowel quality. For example, while the shifts away from /u/ and /æ/ could conceivably be explained on the basis of systematically high estimates of F_1 (i.e., the arrows point in the direction of vowels with higher first formants), those away from /ɑ/ and /e/ are not consistent with this idea. This is not to suggest that formant measurement error does not play a role in accounting for the differences in intelligibility between the NAT and OF signals, but rather, that no simple, systematic difference in formant estimation seems capable of accounting for these differences.

One other possibility worth considering has to do with differences between the natural and synthetic signals during the /h/ and final /d/ intervals. The synthesizer was driven by acoustic measurements during the vowel only, with the /h/ and /d/ segments being generated by some simple rules. As a result, the initial /h/ and, in particular, the final /d/ segments did not always show a very close match between the original and synthetic utterances. It is possible that there is some limited information in the naturally produced consonants that influenced vowel quality. Alternatively, it may be that there was some information in the synthetic consonants that was misleading to listeners in some way. Experiment 2 was designed to test this possibility.

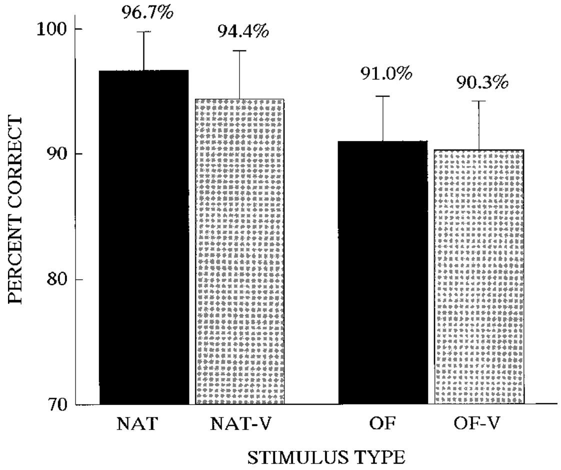

FIG. 7. Percent correct identification for four types of utterances: (a) naturally spoken /hVd/ signals (NAT), (b) the NAT signals with the initial and final consonants edited out (NAT-V), (c) original-formant synthetic /hVd/ signals (OF), and (d) the vowels only from the OF signals (OF-V). Error bars indicate 1 s.d. The percentages above each bar indicate the mean percent correct for each condition.

III. EXPERIMENT 2. METHODS

Experiment 2 presented listeners with four kinds of signals: (a) the 300 natural /hVd/ utterances (NAT); (b) the vowel only from the 300 NAT utterances (NAT-V); (c) the 300 original-formant synthetic /hVd/ utterances (OF); and (d) the vowel only from the 300 OF synthetic utterances (OF-V). The signals were edited from the NAT and OF utterances described above using a simple computer program that was controlled by the hand-measured values of vowel start and vowel end from Hillenbrand *et al.* (1995). After clipping the vowels from the /hVd/ utterances, the NAT-V and OF-V signals were ramped on and off with a 10-ms half-cosine function to prevent onset and offset transients. Listeners consisted of 24 undergraduate students who had taken an introductory phonetics course and had received basic instruction in the use of phonetic symbols for vowels. None of these listeners had participated in experiment 1. Listeners identified each of the 1200 test signals (300 NAT, 300 NAT-V, 300 OF, and 300 OF-V) presented in random order using the same instrumentation and procedures that were described for experiment 1.

IV. EXPERIMENT 2. RESULTS AND DISCUSSION

Overall percent correct values for the four stimulus conditions of experiment 2 are shown in Fig. 7. The main point to be made about Fig. 7 is that both the natural and OF /hVd/ syllables were identified at a slightly higher rate than the corresponding vowel-only utterances. A repeated-measures two-way ANOVA showed significant effects for both factors (natural versus synthetic: $F[1,24]=268.4$, $p<0.01$; syllable versus vowel: $F[1,24]=64.8$, $p<0.01$). As can be seen in Fig. 7, the difference in intelligibility between the /hVd/ and vowel-only conditions is not large overall, and is very small for the OF synthetic stimuli. Newman–Keuls *post hoc* tests showed that this difference was significant only for the naturally spoken stimuli. These results would seem to be consistent with the idea that there is some limited information in the naturally produced consonants that influences vowel quality. If the synthetic consonants were providing misleading information about vowel quality, the results should have shown an improvement in intelligibility for the OF synthetic signals with the removal of the consonants, and little or no change for the natural signals. The small drop in intelligibility that was observed suggests that some very limited information about vowel identity was lost when the natural consonants were clipped off. However, the absolute magnitude of the effect was quite small. The natural vowel-only stimuli remain highly intelligible, and the drop in intelligibility that results from excising the consonants amounts to an average of just 7 additional misidentified vowels out of the 300 that were presented. The primary conclusion from experiment 2, therefore, is that the failure to faithfully model the initial and final consonants can at best explain a very small portion of the difference in intelligibility between the natural and OF synthetic signals.

V. PATTERN RECOGNITION MODELS

Two general conclusions seem likely from the foregoing. First, some changes in listeners' perception (most notably those between the natural and OF stimuli) cannot be readily accounted for by any of the acoustic properties controlled in the experiments. Second, despite this, variations in responses across stimuli are at least partly related to differences in spectral change that were manipulated. The modeling work presented below strives to provide a more detailed assessment of just how far we can go in relating response patterns to selected acoustic properties.

This modeling can be viewed as a way to extend the insights that we were seeking in Fig. 5. There, we measured how average identification rates for each vowel category improved from the FF to the OF condition and we attempted to relate that improvement to average spectral change. The Pearson correlation coefficient ($r = 0.46$) between the x and y coordinates of Fig. 5 gives us a simple index of association between the two quantities. The directness of such an approach is very appealing, and it seems to provide some evidence for the hypothesis being tested. However, it is deficient in several respects. First, it fails to take into account variation among tokens of the same vowel category. Second, a very high correlation should result only under a very limiting assumption: namely, a unit increase in the magnitude of spectral change will result in a uniform change in identification rate, regardless of the direction of the change and of the overall location in formant space of the tokens involved. However, it is easy to imagine cases where this is most unlikely. For example, a token of a vowel whose overall position in formant space is relatively distant from those of its competitors in neighboring categories is likely to be less sensitive to differences in spectral change than a token that is closer to competing tokens.

Nearey and his colleagues have developed pattern recognition methods that can overcome these difficulties (Assmann *et al.*, 1982; Nearey and Assmann, 1986; Andruski and Nearey, 1992; Nearey, 1997). These methods can simultaneously accommodate many aspects of the patterns, including magnitude and direction of spectral change and relative distance from competitors, on a token-by-token basis. We will adopt a variation of the general methodology developed by Nearey and Assmann (1986).

A. Testing and training sets

Much of our modeling uses a two-stage procedure involving the distinction between disjoint training and testing data that is now prevalent in the speech recognition literature. First, a pattern recognition algorithm is constructed based on the statistical regularities in the training data. Second, the parameters of the recognition model are held fixed at the training values, while the algorithm is fed new acoustic data from a testing set that is independent of the training data. This generates predicted response patterns, which are then compared to listeners' performance on the test stimuli.

The training set consisted of 1297 tokens from the larger data of Hillenbrand *et al.* (1995) and will be referred to as the H95 data. The selected tokens included all those with no missing values for any of the measurements required, but excluded the 300 tokens that were used in experiments 1 and 2 above. There was also a corresponding set of responses for each token of the H95 training data. These responses were used for training the logistic regression coefficients in Model B, discussed below. The testing sets consist of the stimuli and responses reported in experiments 1 and 2 above.

B. Stimulus properties and discriminant analysis

We have chosen a representation similar to that of the Canadian studies, using vowel duration, steady-state F_0, and F_1, F_2 and F_3 (all frequencies were log-transformed) at the 20% and 80% time points. (For the flat formant stimuli, the 20% and 80% formant frequencies were equal to the measured steady-state frequencies.) Linear discriminant function analysis of the H95 training data showed that 92.0% of the tokens could be correctly reclassified using these measurements. When the coefficients estimated from the training data were applied to measurements from the natural or OF testing data (with distinct 20% and 80% formant measures), the results were actually higher, 94.0% correct. Thus the chosen measurements are capable of separating the vowel categories relatively well. But to what degree can a recognition algorithm characterize variation in listeners' response patterns?

C. Predicting listener responses on the testing data

1. Model A: A posteriori probabilities from discriminant function analysis of the H95 data

The linear classification functions described above are not only suitable for classification, but they can be used to generate *a posteriori* probabilities (APP scores) of group membership for any given measurement point. APP scores can be viewed as estimates of relative strength of group membership (Nearey and Assmann, 1986). For example, a token whose measurements are near the mean of /i/ and remote from the means of the other categories will have an

APP score for /i/ near unity while scores for the other vowels will be near zero. Similarly, an ambiguous token that is roughly equidistant from /i/, /e/, and /ɪ/ will have a score near 0.33 for each of these vowels, but near zero for the rest.

Such graded membership scores can be compared with confusion matrices, including confusion matrices constructed on a token-by-token basis. Such token-by-token confusion matrices will be referred to below as probability matrices. For each 300 token set, there is both a predicted and an observed probability matrix, each with 300 rows and 12 columns. Predictions based on the linear classification functions from the discriminant analysis of the H95 data will be referred to as Model A predictions.

2. Model B: Predicted probabilities from logistic regression of the H95 perceptual data

Logistic regression provides another technique for generating predicted probability matrices for the testing data (see Nearey, 1989, 1997, for applications of logistic regression to perceptual data). A 12-category polytomous logistic regression was performed using the training measurements as independent variables and the response matrices from the Hillenbrand *et al.* (1995) study as the dependent measures. This method may result in better correspondence to listeners' behavior because, in effect, it can model ambiguity as well as identity. It does this by optimally matching the gradient, probabilistic identification profiles of a group of listeners to each token of the training set, rather than simply predicting nominal correct categories.

D. Measures of association between probability matrices

We will compare predicted and observed probability matrices using methods similar to those of Nearey and Assmann (1986).[5] Three measures of association will be reported. The first is percentage of modal agreement (P_{ma}), defined as the percentage of tokens for which the predicted and observed probabilities show the same modal category, where the modal category is the response with the highest probability for that token. (See the Appendix for a formal definition of P_{ma} and other measures of association used in the pattern recognition work.) Note that this measure does not depend on the nominally correct category of the original recordings. For example, both the listeners and the prediction model might agree that a flat formant token from an original /e/ more closely resembles the /ɪ/ category. These results are shown in Table VI.

The second measure of association, called *correct response correlation* (r_c), is defined as the correlation between predicted and observed probabilities of nominally correct responses to each stimulus (ignoring all incorrect responses). The value of r_c will approach a maximum of 1.0 if and only if *variation* in the relative probabilities of correct identification by listeners is matched by *covariation* in predicted probabilities on a token-by-token basis. These correlations are shown in Table VII. Although these are conventional correlation coefficients from a computational standpoint, it is not clear whether the usual statistical assumptions apply. Therefore, following Nearey and Assmann (1986), we will use nonparametric randomization tests (Edgington, 1980) to assess significance levels of the correlation coefficients.

TABLE VI. Percent modal agreement values (P_{ma}) for each of the stimulus sets used in experiment 1. Columns represent different prediction models. The last column represents an empirical split-sample cross-validation estimate, predicting one-half of the subjects' responses from the other half.

Experiment	Stimulus type	Model A	Model B	Model C	Model C-s	Split-Sample
1	NAT	94.0	97.0	96.3	96.1	99.5
1	FF	71.3	75.3	88.3	87.9	92.5
1	OF	91.7	92.7	94.3	94.2	96.4
2	NAT	94.0	97.0	96.3	96.3	100.0
2	OF	93.7	94.7	96.7	95.6	97.6
2	NAT-V	94.3	96.7	96.0	96.2	100.0
2	OF-V	95.0	95.7	97.0	96.4	97.7

E. Difference correlations

The third measure of association focuses on the ability of the models to predict changes in listener behavior across stimulus conditions. Corresponding stimuli in all the changing formant (i.e., natural and original formant synthetic) stimuli must have exactly the same predicted probabilities because the measurement vectors supplied to the prediction algorithm are identical. However, the measurements for the flat formant synthesis tokens are different, since the formant frequencies are from the steady-state portion. If a pattern recognition algorithm approximates the behavior of our listeners, we would expect *changes* in predicted probabilities of a given token across conditions to be correlated with changes in listeners' responses. Following Nearey and Assmann (1986), we have calculated correlations between changes in predicted probabilities and corresponding changes in observed probabilities. This is done by producing six difference matrices, one for each of the six changing formant conditions (the NAT and OF conditions from experiment 1 plus all four conditions from experiment 2). Each is calculated as the element-by-element difference between the probability matrix of the given changing formant condition and that of the FF condition.

Our analysis here will concentrate on the *correct response difference correlation* (r_{cd}). The calculation is analo-

TABLE VII. Correct response correlations (r_c) for the same analyses as Table VI. Significance levels (by randomization test) shown for completely cross-validated predictions of models A and B only.

Experiment	Stimulus type	Model A	Model B	Model C	Model C-s	Split-Sample
1	NAT	$0.207^{\rm b}$	$0.370^{\rm c}$	0.400	0.345	0.585
1	FF	$0.484^{\rm c}$	$0.547^{\rm c}$	0.850	0.814	0.890
1	OF	$0.399^{\rm c}$	$0.387^{\rm c}$	0.710	0.660	0.814
2	NAT	0.094	$0.305^{\rm c}$	0.249	0.197	0.434
2	OF	$0.408^{\rm c}$	$0.426^{\rm c}$	0.659	0.626	0.812
2	NAT-V	$0.220^{\rm c}$	$0.372^{\rm c}$	0.371	0.326	0.561
2	OF-V	$0.480^{\rm c}$	$0.504^{\rm c}$	0.664	0.605	0.751

$^{\rm a}p<0.01$.
$^{\rm b}p<0.005$.
$^{\rm c}p<0.001$.

TABLE VIII. Difference correct response correlations (r_{cd}) of predicted with observed correct response difference scores when correct response for the flat formant data is subtracted from each of the corresponding tables. Significance levels (by randomization test) shown for completely cross-validated predictions of models A and B only.

Experiment	Stimulus type	Model A	Model B	Model C	Model C-s	Split-Sample
1	NAT	0.364[a]	0.494[a]	0.703	0.669	0.857
1	FF	···		···	···	···
1	OF	0.463[a]	0.541[a]	0.753	0.699	0.788
2	NAT	0.382[a]	0.486[a]	0.694	0.661	0.873
2	OF	0.450[a]	0.554[a]	0.772	0.734	0.835
2	NAT-V	0.356[a]	0.472[a]	0.701	0.667	0.853
2	OF-V	0.395[a]	0.497[a]	0.738	0.698	0.821

[a] $p < 0.001$.

gous to the correct response correlation, except that difference matrices are substituted for the original probability matrices. This correlation will be large only when predicted and observed correct identification rates change in similar ways across conditions. Correct response difference correlations are shown in Table VIII.

F. Benchmark split-sample predictions

As in Andruski and Nearey (1992), we have included benchmark measures of association based on the degree of agreement between subgroups of listeners. This was done by half-sample cross validation. For each experimental condition, subjects were split randomly into two groups. An observed probability matrix was calculated from the responses of approximately half the listeners (10 of 20 for experiment 1, 13 of 25 for experiment 2). This was used to provide nonparametric predictions for the entries of a similar matrix compiled from the remaining data. Measures of association from 200 different random splittings were averaged. These results are presented in the last column of Tables VI–VIII. These figures give a rough estimate of the degree of similarity of empirical response tables when the experiment is repeated with different listeners.

VI. DISCUSSION

A. Changing formant conditions

Consider first the results for the changing formants conditions, i.e., all cases but FF. In Table VI, we see that model A shows modal agreement ranging from 91% to 95% in these conditions. Model B, which had access to gradient aspects of listeners' categorization of the training stimuli, shows even higher agreement (about 93%–97%). A comparison with split-sample benchmark in the last column of Table VI shows that there is still room for improvement: Listeners are somewhat more consistent with each other (modal agreements range from 96% to 100%) than they are with our models.

Much of the similarity of models A and B for all of the non-FF conditions can be attributed to the simple fact that, for the both the listeners and the models, the modal category is the nominally correct category for most of the stimuli. The nominally correct identification rate is 94% for model A and 97% for model B for each of the changing formant conditions. The correct identification rate for listeners varies from about 95% to 100% across conditions (where the ''winning'' category is the one with the plurality of listener votes). However, correct response correlations r_c in Table VII show that more than this overall correspondence of correct responses is involved. Recall that r_c is positive only to the extent that variations in the probability of nominally correct responses covary in predicted and observed tables. Therefore, simply having high average probabilities of correct responses in both observed and predicted matrices will not result in positive correlations. The correlations for model A and model B are all positive and significant for all of the changing-formant conditions. Although the magnitudes of such correlations are modest, we should bear in mind that a ceiling on this correlation is imposed by listener-to-listener variability. An estimate of this ceiling is given in the split-sample column. For the changing-formant cases, the variance accounted for by model A (calculated as the ratio of the squares of the correlation coefficients) is roughly one-third and that by model B is roughly one-half that accounted for in the corresponding split-sample benchmarks.

B. Flat formant condition

In the case of the FF stimuli, neither the *a priori* models (A and B) nor listeners' identifications show nominally correct identification rates nearly as high as in the changing-formant cases. The nominally correct category was chosen by the plurality of listeners in only about 80% of the tokens. Perhaps not surprisingly, the corresponding nominally correct identification rate for model A is considerably lower, only about 60%. Nonetheless, the modal agreement between the two is about 71%. Modal agreement with listeners is higher than the algorithm's correct identification rate because model A predictions showed the same ''modal error'' as listeners in 32 of the 59 tokens nominally misidentified by the plurality of listeners. Model B shows a nominally correct classification rate of about 73%, which is still somewhat lower than listeners. Again, the modal agreement between listeners and model B is higher (about 75%) because model B has also predicted the listeners' ''modal errors'' correctly in 29 of 59 cases. (This agreement on modal errors is slightly less than with model A. The improvement of model B over model A occurs because model B predicts 197 of the 241 correct responses by listeners, while model A correctly predicts only 182 of them.)

While the above results clearly suggest a reasonable degree of correspondence, we also see that listeners are much more consistent with each other than they are with the models. Although the rates of split-sample modal agreement are lower than they were for any of the changing formant conditions, at about 93% they are still more than 20 percentage points higher than the model A results for the FF stimuli.

The general pattern of the modal agreement results is also supported by the correlations between observed and predicted identification rates for nominally correct tokens r_c, given in Table VII. Both model A and B show highly significant correlations. However, the magnitudes of the correlations ($r_c = 0.48$ and 0.54) account for only about one-third of the variance accounted for in the split-sample predictions ($r_c = 0.89$).

While the above models give a reasonable estimation of the overall fit of the predictions, they give only a very indirect view of the relative success of predicting *changes* in categorization *across* conditions. For this, we turn to the correct response difference correlations, r_{cd}, in Table VIII. Recall that this measure involves correlations of the changes (from the FF condition to each of the changing formant conditions) in observed probabilities of correct response with corresponding changes in predicted probabilities. Models A and B both show highly significant correlations for changes in response patterns from the FF condition to each of the other (changing formant) conditions. Model A accounts for only about 17%–35% of the variance accounted for by the split-sample benchmark. Model B fares somewhat better, accounting for about 30%–47% as much as the benchmark. The analysis underlying r_{cd} values is similar in spirit to that of Fig. 5. There are two main differences. First, rather than looking at difference in magnitude of formant change, r_{cd} involves changes in *a posteriori* probabilities (which, for linear discriminant analysis, are closely related to changes in ''relative statistical distance'' to category prototypes) between the two conditions. Second, r_{cd} values are calculated on a token-by-token basis, while Fig. 5 involved averaging over vowel categories.[6] If a similar averaging is done over changes in *a posteriori* probabilities, correlations across vowels are considerably higher for both model A ($r = 0.60$) and model B ($r = 0.71$) than for the spectral distance measure of Fig. 5 ($r = 0.46$).

C. Model C. Predicted probabilities from the experiment 1 perceptual data

Despite the significance of the results reported above, the modest size of the goodness of fit measures for the FF data relative to the split-sample benchmark must give us some pause. However, it should be kept in mind that models A and B were trained only on citation form tokens that must certainly show less variability than the overall population of tokens (produced, e.g., at various speaking rates and stress conditions) to which listeners are exposed. Thus even model B, which was trained on the rather limited degree of gradient behavior in listeners' categorization of the H95 training data, might easily have ''wrapped itself around'' a solution that was dominated by listeners' behavior to relatively prototypical stimuli. It is perhaps not surprising that such predictions might be rather fragile and that they breakdown somewhat when applied to the FF stimuli, which can present rather different stimulus patterns than those in the H95 data for many vowel categories.

We therefore constructed a third, optimized model, model C. Unlike the other two models which are based on the distinct H95 data set, the predictions here are based on observed response probabilities for the 900 stimuli of experiment 1. Model C prediction results are also shown in Tables VI–VIII. This model is included to see how well a model of the same ''size'' as model B might do with the stimuli in experiments 1 and 2. Since it is not a fully cross-validated model, it is likely to present an overly optimistic picture (Efron and Tibshirani, 1993) of prediction errors for new data. However, to the extent that performance of even this model falls below our split-sample benchmarks, we will know that the shortfall is not due simply to the restricted nature of the training data and we will have a useful estimate of the lower bound on how much remains to be explained.

In addition, limited cross validation to distinct listeners and distinct tokens is possible with the available data if we reverse the roles of the training and testing data from those of models A and B. That is, we use logistic regression coefficients ''frozen'' at the values estimated by model C to predict listeners' behavior on the much larger set of natural tokens of the H95 data. This analysis yields a modal agreement of 90.5% with H95 listeners, an r_c correlation of 0.453. If model C is used to classify the H95 training tokens, we find cross-validated classification rate of 90.5%. (This is the same as modal agreement with listeners, because the plurality of listeners' responses in the H95 actually selects the nominally correct category for all stimuli.) This is rather remarkable, given that self-trained linear discriminant analysis on the same data yielded 92.0%. Recall that in model C we are training on measurements based on only 300 different vowels. Those measurements, when optimally mapped to listeners' responses in the three presentation conditions (OF, FF, and NAT), are capable of classifying a completely distinct set of nearly 1300 vowels almost as well as linear discriminant analysis trained on the larger data set itself.

Although we are not able here to provide cross validation to entirely new stimulus tokens in all the conditions of experiments 1 and 2, we can provide true cross validation across different listeners for all conditions of experiment 2 (the last four rows in the model C column of Tables VI–VIII) and we can also provide split-sample cross validation even in the case of experiment 1, by training on the data from one-half of the listeners and testing predictions against the other half (model C-s). In the remaining discussion, we will use the predictions of model C-s, since the measures presented should provide unbiased estimates of prediction success for the same set of stimuli across new groups of listeners. (This is actually the only generalization that the empirical split-sample benchmarks also address. There is no statistical basis for generalizing those results to new stimuli.)

We find that modal agreement numbers for model C-s (Table VI) are uniformly very high, although they are smaller than the split-sample benchmarks by about 1–4 percentage points. Correct response correlations, r_c, in Table VII, are generally within about 0.1 of the corresponding model A and B values, but are higher by about 0.2 to about 0.3 for the FF and the two OF conditions. However, the variance accounted for by model C-s still averages only about half that of the split-sample benchmark, ranging from about 0.21 to 0.83, with the highest value for the FF condition. Correct response difference correlations r_{cd}, in Table VIII, show somewhat more improvement. Model C-s shows values about 0.16–0.31 higher than corresponding entries for models A and B, accounting for about 57%–79% of the variance accounted for by the split-sample benchmark.

VII. GENERAL DISCUSSION

The main purpose of this study was to measure the relative importance of formant frequency movements in the recognition of vowel quality. The substantial difference in overall intelligibility between the OF and FF signals strongly confirms several previous findings indicating that spectral change patterns play a secondary but quite important role in the recognition of vowel quality. The vowels of American and Canadian English (and almost certainly many other vowel systems as well—see Watson and Harrington, 1999, for recent data on Australian English) are more properly modeled not as points in formant space but as trajectories through formant space.[7] However, a simple observation that should not be lost in this discussion of spectral change is that the single-slice spectral measurements reported in studies such as Peterson and Barney (1952) capture most of the information that is needed to represent vowel quality. In the present study, F_0, duration, and steady-state formant measurements were sufficient to signal the intended vowel for roughly three-fourths of the utterances, with nearly all of the misidentifications involving adjacent vowel categories. Strikingly similar identification rates for vowels with static formant patterns were reported by Fairbanks and Grubb (1961), Assmann and Nearey (1986), and Hillenbrand and Gayvert (1993a) in studies using methods that are quite different from those employed here.

The relative importance of formant frequency change varies considerably from one vowel to the next. It was generally the case that the effect of formant flattening was small for vowels that tend to show relatively little formant frequency movement and larger for vowels that tend to show large changes in formant frequencies. The relationship is not quite that simple, however, as demonstrated by the very different effects of formant flattening for /e/, /æ/, and /ɔ/, which showed roughly similar average magnitudes of formant frequency change.

A significant limitation of this study is the exclusive use of the simple /hVd/ environment for all utterances. The relationships between spectral change patterns and vowel identity are guaranteed to be more complex when the consonant environment preceding and following the vowel is allowed to vary. We are currently studying the acoustics and perception of a new multitalker CVC database with variation in both consonants. Preliminary analysis of this database (Hillenbrand and Clark, 1997) using a statistical pattern classifier shows substantially better classification accuracy for two samples of the formant pattern rather than a single sample, in spite of the complexities introduced by variation in consonant environment. The classification advantage for the two-sample case, however, was smaller than we observed in a similar discriminant analysis of our /hVd/ database (Hillenbrand *et al.*, 1995).

The pattern recognition methods described above use a relatively simple approach to vowel specification based on duration, steady-state F_0, and formant frequency measurements at two temporal intervals in the vowel. Such a representation appears to go a considerable distance toward accounting for the results of experiments 1 and 2 since significant correlations were found between various aspects of listeners' behavior and the purely *a priori* predictions of models A and B. However, the agreement between listeners measured by split-sample benchmarks is typically better by a factor of about 2 than even our best predictions. This suggests that we have gone no more than about half the distance to the goal of accounting for listeners' behavior in these experiments.

In our view, the most significant challenge presented by the findings reported here is to explain the difference in intelligibility between the natural signals and the OF synthetic signals. The OF signals were, of course, highly intelligible, indicating that most of the information that is needed to capture vowel identity is preserved by the F_0, duration, and formant measurements that were used to drive the formant synthesizer. But the drop in intelligibility resulting from formant vocoding makes it equally clear that a certain amount of phonetically relevant information was lost.[8] Two possible explanations for this finding were pursued here. First, an analysis of the shifts in vowel identity between the NAT and OF signals suggested that it was unlikely that any simple, systematic error in formant measurement could account for the most common shifts in vowel identity. It was also shown that the failure to faithfully copy the initial and final consonants can at best explain a very small share of this effect. One plausible explanation that was not pursued in this study is that the transformation to a formant representation results in the failure to preserve spectral shape details that are relevant to vowel identity. The formant synthesizer is driven entirely by spectral-peak frequencies, meaning that formant amplitudes, bandwidths, spectral tilt, individual harmonic amplitudes, and other spectral details will often not match well between the natural and synthetic utterances. There has been a fair amount of discussion about the relative contributions of formant frequencies and detailed spectral shape (e.g., Klatt, 1982a,b; Bladon and Lindblom, 1981; Bladon, 1982; Zahorian and Jagharghi, 1986, 1987; Zahorian and Zhang, 1992), but the question is far from resolved. Work that is currently underway involves a close examination of the signals from the present study that showed significant shifts in labeling between the natural and OF formant-synthesis conditions in an effort to understand what specific spectral shape details might play a role in judgments of vowel identity.

ACKNOWLEDGMENTS

This work was supported by a grant from the National Institutes of Health (2-R01-DC01661) to Western Michigan University and by a grant from the Canadian Social Sciences and Humanities Research Council (410-93-0053) to the University of Alberta. We are grateful to Michael Clark, Robert Erickson, and two anonymous reviewers for helpful comments on an earlier draft.

APPENDIX. FORMAL DEFINITIONS OF QUANTITIES USED IN PATTERN RECOGNITION STUDIES

The observed probability matrix has elements $O_s(t,v)$, where the subscript s represents a given stimulus condition (e.g., NAT, FF, etc.). t ranges over the 300 tokens and v over the 12 vowel response categories. Assume the order of the vowels /i,ɪ,e,ɛ,æ,ɑ,ɔ,o,ʊ,u,ʌ,ɜ/. Thus, $O_s(4,3)$ represents the proportion of listeners who responded to stimulus number 4 with vowel category 3. The predicted probability matrix with elements $P_s(t,v)$ is similar in structure, but contains *a posteriori* probability estimates from the pattern recognition models. These matrices are used for the definitions of all other quantities used in the pattern recognition studies.

Percentage modal agreement for a given stimulus condition can be defined as:

$$100\Sigma_t\{M[\operatorname{argmax}_v(O_s(t,v)),\operatorname{argmax}_v(P_s(t,v))]\}/T);$$

where $M(x,y)=1$, if $x=y$ and 0 otherwise, argmax_v indicates the column index of the largest element in a row and T is the total number of tokens (300 in these experiments).

For correct response correlations, define the observed correct probability score for token t as $C_s(t)=O_s(t,c_t)$, where c_t is the column number of the correct response for token t of the observed probability matrix. Thus if stimulus 4 corresponded to the vowel /e/ (vowel number 3) in the original recordings ($s=\text{NAT}$), then $C_s(4)$ is $O_s(4,3)$. Predicted correct probabilities are similarly defined $D_s(t)=P_s(t,c_t)$.

Correct response correlations are then defined as the Pearson correlation between the elements of $C_s(t)$ and the corresponding elements of $D_s(t)$. Randomization tests are calculated by randomizing the token index for the predicted tokens. Correct response difference correlations, r_{cd}, are based on changes in correct identification probabilities between all moving formant conditions and the FF conditions. Define correct response difference scores as $Y_s(t)=C_s(t)-C_{\text{FF}}(t)$ for observed probabilities and as $X_s(t)=D_s(t)-D_{\text{FF}}(t)$ for predicted, where the subscript FF refers to the fixed formant condition. Correct response difference correlations for each condition s (except FF, for which all scores are by definition zero) are then the Pearson correlations between $Y_s(t)$ and $X_s(t)$.

[1] Note that the formant contour for the /æ/ in Fig. 2 shows two relatively stationary segments. While this sort of pattern was not very common overall, it did occur with some regularity for this vowel. There was nothing in the formal procedure for judging steady-state times that would have prevented the research assistants who made these visually based judgments from choosing a frame in the offglide as the steadiest point in the vowel. Inspection of these measurements by the first author showed that this never occurred.

[2] Note that the final /d/ segments of the OF and FF signals matched the original signals fairly closely for overall segment duration, but the original and synthetic signals did not necessarily match with respect to the relative durations of the closure and burst intervals.

[3] Speaker-within-group could also be considered a random effect, although this is not typically done in phonetics experiments. Furthermore, the sampling procedure used did not allow for systematic testing effects as such. Strictly speaking, we have no statistical basis for generalizing to stimuli beyond those used in the experiments. Furthermore, statements about subject and vowel effects are to some degree confounded by individual speaker differences. However, our main focus is on differences among conditions, and we are using ANOVA primarily as a screening tool to draw our attention to possible patterns in the data that are correlated with vowel and speaker group.

[4] As a further check on the stimulus generation process, several dozen of the tokens showing the largest changes in identification from the NAT to OF conditions in experiment 1 were reanalyzed at the University of Alberta using software distinct from that used in Hillenbrand *et al.* (1995). This reanalysis showed that the retracked formant frequencies from the resynthesized stimuli agreed very closely with those of the original signals. Thus, although other aspects of the stimuli (such as formant amplitudes and bandwidths) do vary, we are satisfied that the frequencies of F_1 through F_3 match quite well.

[5] The analyses reported here focus on the probabilities of a single response (either the correct original category or the modal response) in each stimulus row of a probability matrix. We have also performed profile correlation and difference profile correlation analyses similar to those reported in Nearey and Assmann (1986) and Andruski and Nearey (1992) that use all entries of the probability matrices. These analyses revealed patterns generally similar to those reported here.

[6] There is also a third, minor difference. Figure 5 counted the number of tokens whose correct identification changed from OF to FF conditions, regardless of direction, while r_{cd} measures signed changes in probability of correct identification across tokens. A very similar correlation results if the latter difference in response measure is substituted in the analysis of Fig. 5.

[7] An anonymous reviewer suggested that listeners might, ''... base their decisions not on the formant frequencies at the 20%, the 80% or the two combined but on the modal formant frequencies of the vowel production; i.e., the $F1$, $F2$, $F3$ combination that occurs the most often ...'' for the nominally monophthongal vowels. We believe that this is an unlikely possibility. The clearest evidence, in our view, comes from the gating experiments of Nearey and Assmann (1986), showing excellent identification of silent-center vowels (including the nominal monophthongs /ɪ/, /ɛ/, and /æ/) consisting of brief onsets and offsets, but poor labeling of the same segments played in reverse order.

[8] It is worth noting that the method that we used to track changes in vowel color from the natural signals to the OF synthetic versions relies entirely on changes in absolute identification. This method is rather coarse. There were clearly many utterances in which the vowel color appeared to us to change from NAT to OF, but the change was not sufficient to induce a labeling shift for most of the listeners. Conversely, in listening to the natural and OF versions of signals showing a large number of labeling shifts, we have generally been impressed at the subtlety of the change in vowel color.

Ainsworth, W. A. (**1972**). ''Duration as a cue in the recognition of synthetic vowels,'' J. Acoust. Soc. Am. **51**, 648–651.

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. **92**, 390–410.

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

Bennett, D. C. (**1968**). ''Spectral form and duration as cues in the recognition of English and German vowels,'' Lang. & Speech **11**, 65–85.

Black, J. W. (**1949**). ''Natural frequency, duration, and intensity of vowels in reading,'' J. Speech Hear. Dis. **14**, 216–221.

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.

Di Benedetto, M-G. (**1989**). ''Vowel representation: Some observations on temporal and spectral properties of the first formant frequency,'' J. Acoust. Soc. Am. **86**, 55–66.

Edgington, E. (**1980**). *Randomization Tests* (Dekker, New York).

Efron, B., and Tibshirani, R. (**1993**). *Introduction to the Bootstrap* (Chapman and Hall, London).

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

Hillenbrand, J. M., and Clark, M. J. (**1997**). ''Effects of consonant environment on vowel formant patterns,'' J. Acoust. Soc. Am. **102**, 3093(A).

Hillenbrand, J. M., and Gayvert, R. T. (**1993a**). ''Vowel classification based on fundamental frequency and formant frequencies,'' J. Speech Hear. Res. **36**, 674–700.

Hillenbrand, J. M., and Gayvert, R. T. (**1993b**). ''Identification of steady-state vowels synthesized from the Peterson and Barney measurements,'' J. Acoust. Soc. Am. **94**, 668–674.

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

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

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.

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.

Miller, J. D. (**1984**). ‘‘Auditory processing of the acoustic patterns of speech,’’ Arch. Otolaryngol. **110**, 154–159.

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

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. (**1997**). ‘‘Speech perception as pattern recognition,’’ J. Acoust. Soc. Am. **101**, 3241–3254.

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).

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.

Peterson, G., and Lehiste, I. (**1960**). ‘‘Duration of syllable nuclei in English,’’ J. Acoust. Soc. Am. **32**, 693–703.

Peterson, G. E. (**1951**). ‘‘The phonetic value of vowels,’’ Language **27**, 541–553.

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

Rabiner, L. (**1968**). ‘‘Digital formant synthesizer for speech synthesis studies,’’ J. Acoust. Soc. Am. **24**, 175–184.

Shankweiler, D., Strange, W., and Verbrugge, R. (**1977**). ‘‘Speech and the problem of perceptual constancy,’’ in *Perceiving, Acting, and Comprehending: Toward an Ecological Psychology*, edited by R. Shaw and J. Bransford (Lawrence Erlbaum, Hillsdale, NJ).

Stevens (**1959**). ‘‘The role of duration in vowel identification,’’ *Quarterly Progress Report 52*, Research Laboratory of Electronics, MIT.

Stevens, K. N., and House, A. S. (**1963**). ‘‘Perturbation of vowel articulations by consonantal context: An acoustical study,’’ J. Speech Hear. Res. **6**, 111–128.

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

Syrdal, A. K. (**1985**). ‘‘Aspects of a model of the auditory representation of American English vowels,’’ Speech Commun. **4**, 121–135.

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.

Tiffany, W. (**1953**). ‘‘Vowel recognition as a function of duration, frequency modulation and phonetic context,’’ J. Speech Hear. Dis. **18**, 289–301.

Watson, C., and Harrington, J. (**1999**) ‘‘Acoustic evidence of dynamic formant trajectories in Australian English vowels,’’ J. Acoust. Soc. Am. (in press).

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. (**1987**). ‘‘Speaker-independent vowel recognition based on overall spectral shape versus formants,’’ J. Acoust. Soc. Am. Suppl. 1 **82**, S37.

Zahorian, S., and Zhang, Z-J. (**1992**). ‘‘Perception of vowels synthesized from sinusoids that preserve either formant frequencies or global spectral shape,’’ J. Acoust. Soc. Am. Suppl. 1 **92**, S2414(A).