Some effects of duration on vowel recognition

James M. Hillenbrand[a] and Michael J. Clark
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 14 April 2000; accepted for publication 14 September 2000)

This study was designed to examine the role of duration in vowel perception by testing listeners on the identification of CVC syllables generated at different durations. Test signals consisted of synthesized versions of 300 utterances selected from a large, multitalker database of /hVd/ syllables [Hillenbrand *et al.*, J. Acoust. Soc. Am. **97**, 3099–3111 (1995)]. Four versions of each utterance were synthesized: (1) an *original duration* set (vowel duration matched to the original utterance), (2) a *neutral duration* set (duration fixed at 272 ms, the grand mean across all vowels), (3) a *short duration* set (duration fixed at 144 ms, two standard deviations below the mean), and (4) a *long duration* set (duration fixed at 400 ms, two standard deviations above the mean). Experiment 1 used a formant synthesizer, while a second experiment was an exact replication using a sinusoidal synthesis method that represented the original vowel spectrum more precisely than the formant synthesizer. Findings included (1) duration had a small overall effect on vowel identity since the great majority of signals were identified correctly at their original durations and at all three altered durations; (2) despite the relatively small average effect of duration, some vowels, especially /ɑ/-/ɔ/-/ʌ/ and /æ/-/ɛ/, were significantly affected by duration; (3) some vowel contrasts that differ systematically in duration, such as /i/-/ɪ/, /u/-/ʊ/, and /ɪ/-/e/-/ɛ/, were minimally affected by duration; (4) a simple pattern recognition model appears to be capable of accounting for several features of the listening test results, especially the greater influence of duration on some vowels than others; and (5) because a formant synthesizer does an imperfect job of representing the fine details of the original vowel spectrum, results using the formant-synthesized signals led to a slight overestimate of the role of duration in vowel recognition, especially for the shortened vowels. © *2000 Acoustical Society of America.* [S0001-4966(00)03912-6]

PACS numbers: 43.71.An, 43.71.Es [KRK]

I. INTRODUCTION

Duration has long been a key feature in the description and analysis of vowels. The chief phonological question concerns whether duration should be considered a contrastive or redundant feature (Chomsky and Halle, 1968; Roca and Johnson, 1999), and the main phonetic issues have been the measurement of vowel durations under a variety of conditions and the study of duration as a cue in vowel perception. In this study we are neutral regarding the phonological question of whether length should be considered an intrinsic phonological vowel feature in English. Rather we assume the existence of distinct vowel categories that contrast with one another in most phonetic contexts and that are produced with different typical durations in American English. Our focus is on the role played by variations in vowel duration in the recognition of vowel identity. Specifically, we studied the perception of 300 /hVd/ syllables that were synthesized in four different ways: (1) an *original duration* condition in which the duration of each vowel was matched as closely as possible to that of the original utterance, (2) a *neutral duration* condition in which the synthesis control parameters were linearly stretched or contracted to produce a fixed vowel duration of 272 ms (the mean of all 300 utterances), (3) a *short duration* condition in which vowel duration was fixed at 144 ms (two standard deviations below the mean), and (4) a *long duration* condition in which vowel duration was fixed at 400 ms (two standard deviations above the mean).

A. Measurement studies

The central phonetic fact underlying this study is the well-known observation that American English vowels differ from one another in average duration. Of particular interest are the many pairs of spectrally similar vowels that differ in duration, pairs such as /i/-/ɪ/, /u/-/ʊ/, /æ/-/ɛ/, /e/-/ɛ/, /ɑ/-/ʌ/, and /ɔ/-/ɑ/. Average vowel duration measurements for the 12 vowel types used in the present experiment are summarized in Fig. 1. The data from Crystal and House (1988) and van Santen (1992) are from connected speech, while the Hillenbrand *et al.* (1995) and Black (1949) measurements are from CVC syllables. There is, of course, quite a bit of variability in the absolute durations associated with each vowel type across the four studies, reflecting the differences in speech material. As expected, the two connected speech studies show shorter average durations than the two studies using citation-form syllables. The longer durations in Hillenbrand *et al.* than in Black are related primarily to the use of a final voiced stop (/hVd/) in Hillenbrand *et al.* as compared to a final voiceless stop (/tVp/) in Black (House and Fairbanks,

[a]Electronic mail: james.hillenbrand@wmich.edu

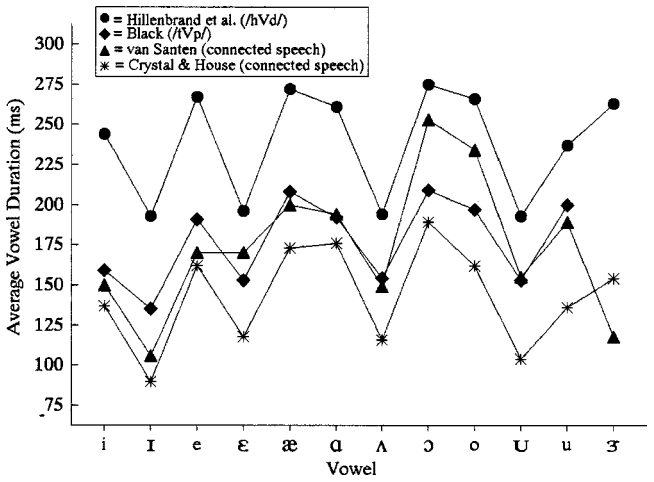

FIG. 1. Average vowel durations from four studies: Hillenbrand *et al.* (1995), Black (1949), van Santen (1992), and Crystal and House (1988). Measurements for /ɝ/ are not available from Black.

1953). Despite these differences in absolute duration, however, the four studies show rather similar patterns of duration differences across the vowel categories. Correlations among the six possible pairings of the four functions shown in Fig. 1 ranged from 0.70 to 0.95, with an average of 0.84. As will be seen later, the results of our synthesis experiments show that listeners are tacitly aware of these differences in typical duration and make some use of this knowledge in making judgments about vowel identity.

B. Pattern recognition studies

The role of duration in vowel identification has been studied indirectly through the use of pattern recognition experiments. In work of this type, a statistically based pattern classifier is used to determine the separability of vowels based on various combinations of acoustic measurements. For example, Zahorian and Jagharghi (1993) used a discriminant classifier to identify signals in a database of 2922 CVC syllables formed from nine initial consonants, 11 vowels, and eight final consonants. Zahorian and Jagharghi's main interest was the comparison of two methods of representing the spectral characteristics of vowels, but they also tested the value of duration for improving vowel categorization. Zahorian and Jagharghi reported a statistically nonsignificant improvement in classification accuracy of less than 1% when duration was added to the spectrally based acoustic measurements that were used to train the pattern classifier, suggesting a very limited role for duration in vowel separability. Very different conclusions were reached by Hillenbrand *et al.* (1995), who used a discriminant classifier to identify signals in a database of /hVd/ syllables produced by men, women, and children (12 vowels$\times$139 talkers$=$1668 syllables). Results showed that inclusion of duration in the parameter set resulted in consistent improvements in category separability, especially for the simpler parameter sets involving very few spectrum-related variables (e.g., a single sampling of F_1 and F_2 at steady state). Similar findings were reported by Hillenbrand *et al.* (2000) in a discriminant analysis study of CVC syllables formed from seven initial consonants, eight vowels, and six final consonants spoken by 12 talkers. Consistent with Hillenbrand *et al.* (1995), the results showed a modest but consistent improvement in classification accuracy with the addition of duration measures. Particularly large improvements in category separability were seen for /æ/ and /ɛ/, but there were also substantial improvements for /ɪ/ and /ɑ/. Finally, a study of Australian English vowels by Watson and Harrington (1999) showed a small improvement in classification accuracy when measurements of formant trajectories were augmented by duration measures.

C. Perception studies

More direct evidence on the role of duration in vowel identification comes from a series of perception experiments using synthetic speech or modified natural speech. For example, Tiffany (1953) recorded 12 vowels spoken by four phonetically trained men under a variety of conditions differing by pitch and phonetic context (e.g., /tVp/ syllables versus vowels in isolation). The talkers also produced long sustained vowels, from which segments of different durations were clipped (80 ms, 200 ms, 500 ms, and 8 s). The signals were identified by listeners who had some training in phonetics. Some duration effects emerged from the listening data; for example, some vowels with long typical durations (e.g., /e/ and /ɑ/) were more likely to be correctly identified at longer durations, while others with short typical durations (e.g., /ɪ/ and /ʊ/) were better identified at the shorter durations.

Stevens (1959) synthesized /dVs/ syllables at durations ranging from 25 to 400 ms, one series with front vowels (to be identified by listeners as /i, ɪ, ɛ, æ/) and another with back vowels (to be identified as /u, ʊ, ʌ, ɑ/). Several effects were observed which are consistent with the idea that listeners use both spectrum and duration in identifying vowels. For example, for durations less than about 100 ms, vowels with formant specifications appropriate for /æ/ were judged to be /ɛ/. Similarly for vowels with formant specifications appropriate for /ɑ/, the stimuli shorter than 100 ms were judged as /ʌ/. Less robust shifts were seen from /i/ to /ɪ/ and from /u/ to /ʊ/, and these tended to occur only for extremely short vowels. In a related study, Ainsworth (1972) synthesized two-formant vowels with formant values covering the English vowel space with durations ranging from 120 to 600 ms. Listeners were influenced in a manner generally consistent with observed durational differences among vowels. For example, signals with F_1 and F_2 values generally similar to those found for /u/ and /ʊ/ were more likely to be identified as /ʊ/ if short and /u/ if long.

A synthesis experiment by Huang (1986) yielded somewhat equivocal results. Listeners were presented with nine-step synthetic continua contrasting a variety of spectrally similar vowel pairs at durations of 40, 90, 140, and 235 ms. While the expected duration-dependent boundary shifts occurred (e.g., the /i/-/ɪ/ boundary shifted in the direction of /ɪ/ at shorter durations), duration differences much larger than those observed in natural speech were typically needed to move the boundaries. For duration differences approximating those found in natural speech, boundary shifts were small or

nonexistent. Huang also reported unexpected boundary shifts for lax-lax pairs such as /ʊ/-/ʌ/ that do not differ in duration.

The Stevens (1959), Ainsworth (1972), and Huang (1986) studies, which used synthetically generated vowels with static formant patterns, should be interpreted with some caution since it is well known that vowel color tends to be considerably more ambiguous for signals with static formant patterns than for vowels showing natural patterns of spectral change over time (e.g., Hillenbrand and Gayvert, 1993; Hillenbrand and Nearey, 1999; Fairbanks and Grubb, 1961). Consequently, it is possible that studies using static synthetic vowels have overestimated the importance of duration information.

Daniloff *et al.* (1968) measured the intelligibility of naturally produced vowels in /hVd/ syllables under several conditions of time and frequency distortion. Intelligibility was found to be more vulnerable to frequency division than time compression. Except at the most extreme degrees of time compression, syllables were correctly identified as to vowel category at 90% or better. As expected, time compression affected long vowels (/æ, ɑ, ɔ, e/ much more than short vowels (/ɪ, ɛ, ʌ, ʊ/), with intermediate effects for medium duration vowels (/ɝ, u, i/). As expected, most errors in the Daniloff *et al.* data involved the misidentification of longer vowels as their shorter-duration neighbors.

Mixed results on the effects of duration on vowel identity were reported by Strange *et al.* (1983) in experiments using silent-center stimuli—signals consisting of brief onglides and offglides, with the center vowel nuclei replaced by a variable-duration silent gap. Listeners were presented with three kinds of silent-center stimuli: (1) durational information retained (i.e., onglides and offglides separated by an amount of silence equal to the duration of the deleted vowel nucleus), (2) durational information neutralized by setting the silent intervals for all stimuli equal to the shortest vowel nucleus, and (3) durational information neutralized by setting the silent intervals for all stimuli equal to the longest vowel nucleus. Results were mixed: shortening the silent interval to match the shortest vowels did not increase error rates relative to the natural duration condition, but lengthening the intervals to match the longest vowels produced a significant increase in error rates. The authors speculated that the results for the lengthened signals may have been ''... due to the disruption of the integrity of the syllables, rather than misinformation about vowel length; that is, subjects may not have perceived a single syllable with a silent gap in it, but instead, heard the initial and final portions as two discrete utterances'' (Strange, 1989, p. 2140). While the experiments using silent-center and related stimuli (see also Nearey and Assmann, 1986) have clearly been quite important, the uncertainty that results from the kind of interpretive problem described by Strange represents an important weakness of this class of experiments. The present experiments adopt an approach that is generally similar to that of Strange *et al.*, but we will attempt to address this limitation by using stimuli modeled on naturally spoken CVC syllables rather than silent-center stimuli.

D. Summary

The picture that emerges from the pattern recognition and perception studies described earlier is not entirely clear. Much of the evidence is consistent with the idea duration plays a modest but measurable role in vowel recognition, but the findings are far from uniform. The present study was designed to explore this question further by testing listeners on the identification of resynthesized /hVd/ utterances under four duration conditions. We were especially interested in conducting a relatively large-scale study in which the spectral properties of the test signals, and especially the patterns of formant frequency change over time, were modeled as closely as possible on naturally spoken speech signals.

II. EXPERIMENT 1

A. Methods

1. Test signals

The test signals consisted of four different synthesized versions of 300 /hVd/ utterances that were sampled from the 1668 utterances recorded by Hillenbrand *et al.* (1995). The full database consisted of recordings of 12 vowels (/i,ɪ,e,ɛ,æ,ɑ,ɔ,o,ʊ,u,ʌ,ɝ/) in /hVd/ syllables spoken by 45 men, 48 women, and 46 10- to 12-year-old children. The 300-stimulus subset was selected at random from the full database, 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. This same 300-syllable subset had been used in an earlier study of the effects of formant contour on vowel recognition (Hillenbrand and Nearey, 1999).

2. Acoustic measurements

Acoustic measurements of the /hVd/ syllables consisted of formant contours for F_1-F_4 measured from LPC spectra (sampled every 8 ms) and edited by hand during the vowel using methods that are described in detail in Hillenbrand *et al.* (1995). Measurements were also made of (a) the F_0 contour (also edited by hand), (b) the onset of the vowel, and (c) the offset of the vowel. Vowel onsets and offsets were judged by visual inspection using standard measurement criteria (Peterson and Lehiste, 1960). The distribution of vowel durations in the 300-stimulus subset was approximately symmetrical with a mean duration of 274.0 ms, a median of 268.5 ms, and a standard deviation of 65.3 ms.

3. Synthesis method

The Klatt and Klatt (1990) formant synthesizer, running at a 16-kHz sample rate, was used to generate four sets of synthetic signals differing in vowel duration (see Fig. 2). An *original duration* (OD) set was generated in a straightforward way from the measured F_0 and formant contours, so the vowel durations of these signals matched those of the

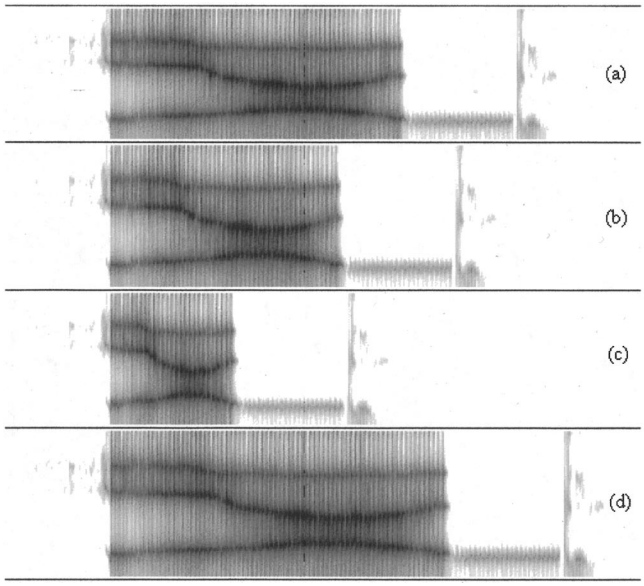

FIG. 2. Spectrograms of the four types of synthetic signals used in experiment 1: (a) original duration (OD), (b) neutral duration (ND), (c) short duration (SD), and (d) long duration (LD). The original signal was /hæd/ spoken by a child.

original signals, within the limits of measurement error and the 8-ms frame rate. Procedures for synthesizing initial /h/ and final /d/ segments for these signals are described in detail in Hillenbrand and Nearey (1999; hereafter HN99). Briefly, the initial /h/ was synthesized by (a) setting the frequencies of all formants to their values measured at vowel onset, (b) setting voicing amplitude to zero and aspiration amplitude to the measured rms intensity of the signal being synthesized, and (c) setting the bandwidth of F_1 to 300 Hz. A final /d/ was simulated by (a) ramping F_1 100 Hz below its measured value at the end of the vowel in four 25-Hz steps, and (b) switching from the cascade to the parallel branch of the synthesizer and setting the resonator gains of F_2–F_6 30 dB below the F_1 resonator gain, producing a ''voice bar'' with energy primarily at F_1. Since we were not 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. Formant frequencies for F_1–F_3 and fundamental frequency during the vowel were set to the original measured values. Formant amplitudes during the /h/ and vowel were set automatically by running the synthesizer in series mode, formant bandwidths were kept at their default values (see HN99), the frequency of F_4 was set separately for each vowel and talker group based on data from Hillenbrand *et al.* (1995), and the frequencies of F_5 and F_6 were based on Rabiner (1968).

The synthesis parameter files for the OD signals [Fig. 2(a)] served as the basis for generating parameter files for (a) a *neutral duration* (ND) set with vowel durations fixed at 272 ms, the grand mean of all vowel durations from Hillenbrand *et al.* (1995), rounded to nearest 8-ms frame; (b) a *short duration* (SD) set with vowel durations fixed at 144 ms, 2 standard deviations below the grand mean (again rounded to closest 8-ms frame); and (c) a *long duration* (LD) set with vowel durations fixed at 400 ms, 2 standard deviations above the grand mean. The parameter files for the ND, SD, and LD signals were created from the OD parameter files simply by resampling the contours of F_0–F_3 during the vowel using linear interpolation. Parameter settings during the /h/ and /d/ segments were unchanged. Examples of ND, SD, and LD signals are shown in panels (b)–(d) of Fig. 2.

TABLE I. Overall recognition rates for the four duration conditions of experiments 1 and 2 (OD=original duration, ND=neutral duration, SD=short duration, LD=long duration). Standard deviations are shown in parentheses.

Duration condition	Experiment 1	Experiment 2
OD	91.7(3.3)	96.0(2.7)
ND	90.4(3.5)	94.1(2.9)
SD	82.4(4.9)	91.4(5.2)
LD	89.5(2.2)	90.9(3.3)

4. Listening test

Fifteen phonetically trained subjects served as listeners. Twelve of the listeners were graduate students in the speech-language pathology program at Western Michigan University and the other three were faculty members in the same department. Listeners with training in phonetic transcription were chosen because of the findings of Assmann *et al.* (1982) showing that many apparent identification errors made by untrained subjects are, in fact, simply errors in transcription. The regional dialect characteristics of the listeners were a fairly close match close to those of the talkers. Dialect was assessed by a trained phonetician using an interview procedure similar to that described in Hillenbrand *et al.* (1995). Eight of the listeners were raised in southwest Michigan, three in Chicago, two in the northeast (Massachusetts and New Jersey), and one each in Iowa and California. Listeners were tested one at a time in a quiet room in two sessions lasting about 35–40 min. Listeners identified each of the 1200 test signals (300 OD, 300 ND, 300 SD, and 300 LD) presented unblocked in a single 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.

B. Results and discussion

Overall recognition rates for the four duration conditions are shown in Table I (experiment 2 results, shown to the right in Table I, will be discussed later). The 91.7% recognition rate for the OD signals is slightly higher than the 89.8% rate for the same set of signals that were used in an earlier study (averaged across the two ''OF'' conditions from HN99). More importantly, the recognition rate for the OD

signals is lower than the 94.5% rate for the original, naturally produced signals from the full 1668-signal database (Hillenbrand *et al.*, 1995). It is also lower than the 96.0% recognition rate for naturally spoken versions of the 300-utterance subset used in HN99. As will be discussed in greater detail later, the slightly lower recognition rate for the OD synthetic signals as compared to the natural signals is due to a small but important limitation of the formant vocoding method to faithfully model some perceptually relevant details of the original vowel spectrum.

It can be seen that the OD signals were the most intelligible, followed by the ND, LD, and SD signals. The only numerically large effect, however, is the drop in intelligibility that occurred as a result of vowel shortening, with the SD signals being nearly 10 percentage points less intelligible than the OD signals. A two-way repeated measures analysis of variance showed a highly significant effect for duration condition $[F(3,42)=38.3,\ p<0.0001]$ and vowel $[F(11,154)=22.2,\ p<0.0001]$ and a significant duration by vowel interaction $[F(33,462)=19.6,\ p<0.0001]$. Bonferroni *post hoc* tests showed significant differences among all pairs of duration conditions with the exception of ND versus LD.

A detailed analysis of the specific changes in vowel identity that occurred in the three duration-modification conditions will not be undertaken here. As will be explained below, there are some key aspects of the findings of experiment 1 that do not replicate when a synthesis method is used that more faithfully models the detailed spectrum of the original vowels. Briefly, the vowels that were most affected by shortening were /æ/, which tended to shift to /ɛ/, and /ɑ/, which tended to shift to /ʌ/. Similarly, but to a lesser extent, the opposite shifts in vowel identity tended to occur as the most common effects of vowel lengthening; i.e., lengthened /ɛ/ tended to shift toward /æ/ and lengthened /ʌ/ tended to shift to /ɑ/ or /ɔ/. Considerably less common were duration-induced shifts in vowel identity affecting the /i/-/ɪ/ contrast, the /u/-/ʊ/ contrast, or distinctions among the cluster /ɪ/-/e/-/ɛ/. A more detailed discussion of the effects of duration on the recognition of individual vowels will await the outcome of experiment 2.

In summary, experiment 1 showed that the effect of altering vowel duration was modest overall since the great majority of signals were accurately identified at their original durations and with vowel duration set to a neutral value, shortened, and lengthened. There were some significant effects, however, including (1) vowels with long typical durations, especially /æ/ and /ɑ/, tended to shift to adjacent vowels with shorter typical durations when shortened, (2) vowels with short typical durations, especially /ʌ/ and /ɛ/, tended to shift to adjacent vowels with longer typical durations, and (3) vowel shortening had a considerably greater effect on vowel identity than vowel lengthening.

III. EXPERIMENT 2

The purpose of experiment 2 was to determine the generality of the effects observed in experiment 1 by conducting a similar experiment but using a very different method to synthesize the test signals. The logic that is implicit in experiment 1 is that the spectral properties associated with vowels are held constant across the four duration conditions—*and matched as closely as possible to the original signals*—while vowel duration is varied. A problem that is inherent in this approach is that the spectral properties of vowels are imperfectly represented by the formant synthesis method. The difference in intelligibility between the original signals and the formant synthesized versions of those same signals is not especially large, but it is quite real. For example, in HN99, the recognition rate for the 300 naturally produced /hVd/ signals averaged 6.3 percentage points higher than that of the formant-synthesized versions of the same signals, which are identical to the OD signals used in experiment 1. As a consequence, it might be argued that experiment 1, by failing to faithfully model the perceptually relevant spectral cues to vowel identity, may have overestimated the importance of duration in vowel recognition. Experiment 2 was an attempt to explore this possibility by directly replicating experiment 1 using a synthesis method that more accurately models the spectral characteristics of the original signals. The experiment used the same 300-stimulus database and the same four duration conditions (OD, ND, SD, and LD). However, a synthesizer based on the summation of sinusoidal Fourier components was used to generate the test signals. As will be explained below, this synthesizer does a significantly better job of coding the spectral properties of vowels, resulting in a set of OD signals whose intelligibility is essentially indistinguishable from that of the natural produced signals upon which they are based.

A. Methods

1. Test signals

Four sets of synthetic test signals were generated with a sinusoidal synthesizer that has a number of features in common with the method described by McAuley and Quatieri (1986). As in experiment 1, we generated OD, ND, SD, and LD versions of each of the 300 /hVd/ syllables. Briefly, the sinusoidal synthesizer can be thought of as something akin to resynthesis using an inverse Fourier transform, with the important exceptions that (1) phase relations among spectral components are not preserved,[1] and (2) sinusoidal components are generated only for spectral peaks (i.e., harmonic peaks in voiced regions and harmonically unrelated spectral peaks in unvoiced regions, but not nonpeak Fourier components). The analysis begins with the calculation of a high-resolution Fourier spectrum over a relatively large hamming-windowed segment of the speech signal. A 64-ms window was used in this experiment. Spectral peak frequencies and amplitudes are then measured from the narrow-band spectrum. These peaks will correspond primarily to voice-source harmonics during voiced intervals, but the analysis proceeds in the same way for both voiced and unvoiced intervals. The analysis window is then advanced by some constant (8 ms in the present case), and the measurement of spectral peaks continues to the end of the signal. For each spectral peak that is measured from the Fourier spectrum, a sinusoid is generated at the measured frequency and amplitude and with a duration equal to the frame rate (8 ms in our implementation). Since it is essential that phase discontinuities not occur

at the boundaries between frames, it is necessary to track spectral peaks from one frame to the next in much the same way that envelope peaks are tracked from frame to frame in a formant tracker. If the tracking algorithm determines that a given spectral peak is continuous from one frame to the next, the frequency and amplitude of the peak are linearly interpolated through the frame. Further, the starting phase of the sinusoid in frame $n+1$ is adjusted to be continuous with the ending phase of the sinusoid in frame n. If the tracking algorithm determines that a spectral peak in frame n does not continue into frame $n+1$, the amplitude of the sinusoid is ramped down to zero. Similarly, if a spectral peak is found in a given analysis frame that is determined to be new (i.e., not continuous with a peak in the previous frame), the amplitude of the sinusoid is ramped up to its measured amplitude.

One of the many elegant aspects of the sinusoidal approach to synthesis is that the manipulation of speech rate, either globally or frame by frame, is exceedingly simple. Altering duration is simply a matter of changing the durations of the individual sinusoids from the 8-ms frame rate (or whatever the frame rate happens to be) to some other value. For example, for the frame rate used here, decreasing duration by a factor of 2 is simply a matter of changing the durations of the individual sinusoids from 8 to 4 ms. Similarly, increasing duration by a factor of 1.5, for example, is accomplished by changing the durations of the individual sinusoids from 8 to 12 ms. With this rate-manipulation method, the evolution of spectral shape from one frame to the next remains constant from one duration condition to another, and what varies is the rate at which one spectral shape evolves into the next. This is also true of the formant synthesis method that was used in experiment 1. The most important difference between the two methods, in our view, is in the degree to which the detailed spectral shape of the original signal is preserved. With a formant synthesizer the match between the spectral shape of the original signal and that of the resynthesized signal is only approximate since only the frequencies of broad envelope peaks are preserved (and even then, only within the limits of formant estimation, which is imperfect—see HN99 for a discussion). The match is much better with the sinusoidal method since many more spectral details—*all narrow-band spectral peaks*—are preserved in the resynthesis (see Fig. 3).

In generating a set of OD signals comparable to the set used in experiment 1, the default 8-ms sinusoidal duration was simply left unmodified. For the ND set, this duration was adjusted on a signal-by-signal basis to a value that was sufficient to produce a vowel duration of 272 ms for all signals. Sinusoidal durations were modified only during the vowel and not during the /h/ and /d/ segments of the signal. The SD and LD sets were generated using the same method, but with vowel durations fixed at 144 and 400 ms, respectively.

2. Subjects and procedures

A separate group of 14 listeners participated in experiment 2. As in experiment 1, the listeners were phonetically trained and, based on a dialect interview, were judged to speak a dialect that was quite similar to that of the speakers. Instrumentation and experimental procedures were identical to experiment 1.

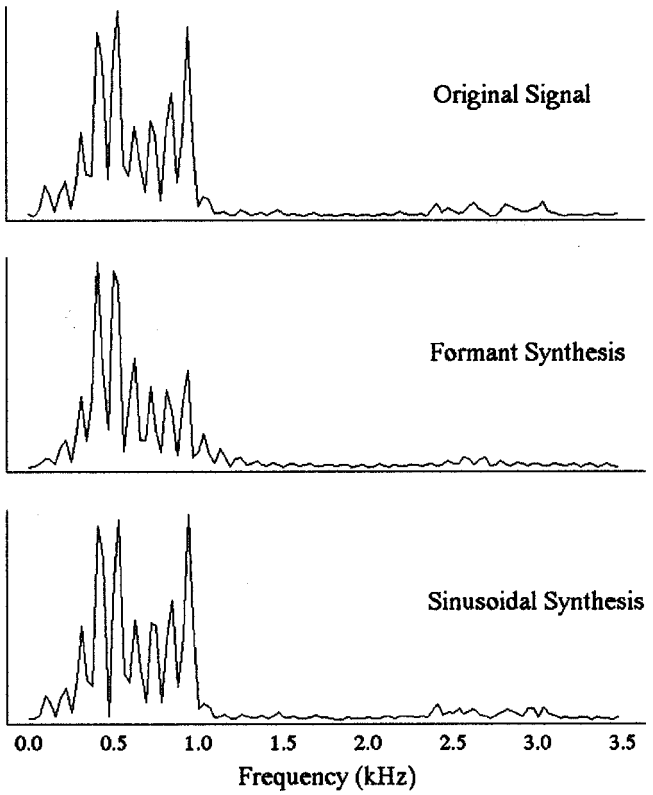

FIG. 3. Fourier spectra computed at roughly the center of the vowel /o/ from (a) the original speech signal, (b) the OD formant synthesized version of the same signal, and (c) the OD sinusoidal synthesized version of the same signal. Note that the match in spectral detail is much closer for the sinusoidal synthesis than for formant synthesis.

B. Results and discussion

1. Effects of synthesis method

An assumption underlying the design of experiment 2 was that the sinusoidal synthesizer would preserve the detailed spectral properties of vowels better than the formant synthesizer. We speculated that this may have led to a slight overestimate of the relative importance of duration in vowel recognition in experiment 1. Accordingly, we begin our analysis of experiment 2 by comparing the recognition of the OD signals from experiments 1 and 2. Figure 4 shows average recognition rates by vowel for OD signals produced with the formant synthesizer versus the sinusoidal synthesizer. It can be seen that the sinusoidal OD signals were more intelligible than the formant synthesized versions (96.0% versus 91.7%). The 96.0% recognition rate for the sinusoidal OD signals is virtually identical to the recognition rate reported in HN99 for naturally spoken versions of the same signals.[2] A two-way analysis of variance for synthesis method and vowel, with repeated measures on the vowel factor, showed significant effects for both factors, as well as a significant interaction [synthesis method: $F(1,27)=32.1$, $p<0.0001$; vowel: $F(11,27)=15.3$, $p<0.0001$; method by vowel: $F(11,297)=6.4$, $p<0.0001$]. The nature of the interaction is readily apparent in Fig. 4, which shows considerable vari-

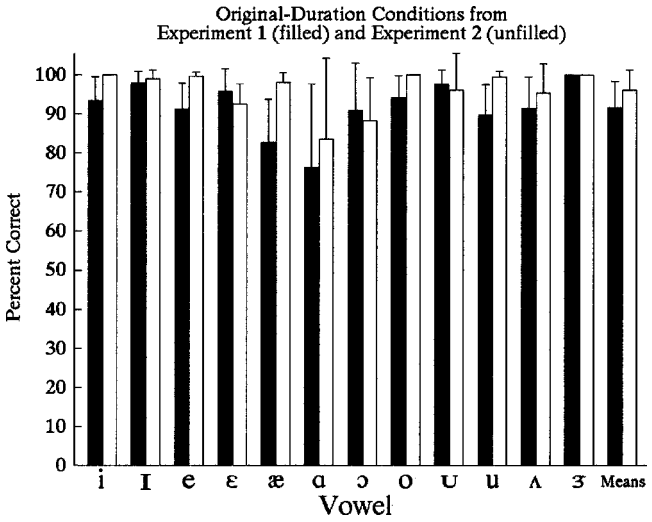

FIG. 4. Percent correct for the original duration conditions from experiment 1 (filled) and experiment 2 (unfilled). Error bars show one standard deviation.

ability across the 12 vowels in the synthesis method effect. As expected, most of the vowels show higher recognition rates when generated with the sinusoidal synthesizer than the formant synthesizer, although the advantage is much larger for some vowels (e.g., /æ/, /u/, and /e/) than others (e.g., /ɪ/ and /ʌ/). Note also that /ɛ/, /ɔ/, and /ʊ/ were actually slightly *more* intelligible in the formant synthesis condition. The findings for these three vowels are consistent with results from HN99's comparison of formant synthesized and natural speech, which showed *tendencies* for (1) signals with formant values in the /æ/-/ɛ/ region to be heard as /ɛ/ when formant synthesized, (2) signals with formant values in the /ɑ/-/ɔ/ region to be heard as /ɔ/ when formant synthesized, and (3) signals with formant values in the /u/-/ʊ/ region to be heard as /ʊ/ when formant synthesized (see especially Fig. 6 of HN99).

2. Effects of duration

Average recognition rates and standard deviations for the four duration conditions of experiment 2 are shown in the right-most columns of Table I. Relative to the original duration signals, there is a modest drop in intelligibility of 1.9% for the neutral duration signals and drops of 4.6% and 5.1% for the shortened and lengthened signals, respectively. A two-way repeated measures analysis of variance showed highly significant effects for duration condition [$F(3,39) = 12.4$, $p<0.0001$] and vowel [$F(11,143)=16.8$, $p<0.0001$] and a significant duration by vowel interaction [$F(33,429)=19.0$, $p<0.0001$]. Bonferroni *post hoc* tests showed significant differences among all pairs of duration conditions except ND-SD and SD-LD. The duration effects observed here differ in some important respects from those of experiment 1. A major finding of experiment 1 was a substantial asymmetry in the effects of shortening versus lengthening, with shortening resulting in a much larger drop in vowel intelligibility than lengthening. There was no evidence for this asymmetry in experiment 2: the average effect of vowel shortening in experiment 2 was considerably smaller than in experiment 1, and was not significantly greater than the average effect of vowel lengthening.

TABLE II. The most frequent changes in vowel identity resulting from vowel shortening or vowel lengthening in experiment 2. The percentages in the column to the right reflect the number of shifts in vowel identity (OD→SD or OD→LD) divided by the number of opportunities for such shifts to occur.

Vowel shift	Percentage
Effects of vowel shortening	
/ɔ/→/ɑ/ or /ʌ/	43.0
/æ/→/ɛ/	20.7
/ɑ/→/ʌ/	9.4
Effects of vowel lengthening	
/ʌ/→/ɑ/ or /ɔ/	36.0
/ɛ/→/æ/	18.9

Table II provides a summary of the most frequent changes in vowel identity that occurred as a result of shortening and lengthening. The shifts in vowel identity that are shown in this table are based on confusion matrices which focused on instances in which a given listener identified the OD version of a signal correctly (i.e., as the vowel intended by the talker) but the duration-modified version of the same utterance incorrectly. For example, suppose that listener 1 correctly identified an OD signal that was intended as /æ/ by the talker but that same listener identified the SD version of the same stimulus as /ɛ/. The count in the cell associated with row /æ/ and column /ɛ/ in the confusion matrix would be incremented by 1. To simplify the presentation, the full confusion matrices are not shown, and only the most frequency shifts in vowel identity are listed in Table II. The percentages that are shown to the right are based on the number of shifts in vowel identity divided by the number of opportunities for such shifts to occur. As the table shows, only /ɔ/ was strongly affected by shortening, with 43.0% of the tokens shifting to /ɑ/ or /ʌ/. The only other vowel shifts of any consequence consisted of the roughly one-fifth of the OD tokens of /æ/ that were heard as /ɛ/ when shortened, and a modest number of /ɑ/ tokens that shifted to /ʌ/. The effects of vowel lengthening, shown in the bottom half of Table II, are nearly the mirror image, with lengthened /ʌ/ tending to shift to /ɑ/ or /ɔ/, and /ɛ/ tending to shift to /æ/. Conspicuous by their absence in Table II are several shifts in vowel identity which might have been expected based on observed durational differences among vowels but which occurred rarely or not at all. For example, there were very few cases of shortened /i/ shifting to /ɪ/ (0.6%), no cases of shortened /u/ shifting to /ʊ/, and few cases of shortened /e/ shifting to /ɪ/ or /ɛ/ (2.2%). Similarly, lengthened /ɪ/ seldom shifted to /i/ (0.6%), lengthened /ɛ/ or /ɪ/ never shifted to /e/, and lengthened /ʊ/ seldom shifted to /u/ (1.3%).

In summary, experiment 2 showed the following: (1) duration has a measurable but rather modest overall effect on vowel perception since the overwhelming majority of the signals were identified correctly at their original durations and at all three altered durations, (2) vowel shortening and vowel lengthening produced statistically equivalent reductions in vowel intelligibility of 4.5%–5.0%, (3) the vowels

that are most affected by duration are the /ɑ/-/ɔ/-/ʌ/ cluster and the /æ/-/ɛ/ pair, (4) in spite of consistent differences in average durations, vowels that are hardly affected at all by duration are the /i/-/ɪ/ and /u/-/ʊ/ pairs, and the /ɪ/-/e/-/ɛ/ cluster. The similarities and differences between experiments 1 and 2 will be treated in Sec. V.

IV. PATTERN RECOGNITION TESTS

The final question that we wish to consider has to do with the finding that some pairs and clusters of vowels (/ɑ/-/ɔ/-/ʌ/ and /æ/-/ɛ/) were influenced fairly strongly by the modification of duration, while others (/i/-/ɪ/, /u/-/ʊ/, and /ɪ/-/e/-/ɛ/) were minimally affected by duration modification. There is nothing obvious about the magnitude of the durational differences among the vowels that can readily explain these variations in the influence of duration on vowel recognition. For example, based on the Crystal and House (1988) data, /æ/ averages about 18% longer than /ɛ/, while /i/ averages about 41% longer than /I/. However, it is the /æ/-/ɛ/ pair which shows a robust duration effect, while shortening /i/ or lengthening /ɪ/ have very little effect on vowel identity. Similarly, while /u/ is about 51% longer on average than /ʊ/, and /ɔ/ is about 30% longer than /ɑ/, it was the contrast between /ɑ/ and /ɔ/ that was affected by duration and not the contrast between /u/ and /ʊ/. We believe that there is a fairly straightforward explanation for these apparently contradictory findings. A combination of listening to the /hVd/ signals and close examination of the acoustic measurements led us to speculate that non-duration-sensitive pairs such as /i/-/ɪ/ and /u/-/ʊ/ are quite distinct from one another based on their spectral properties (F_0 and formant trajectories) and, as a consequence, less dependent on duration for their separation. On the other hand, vowels such as /ɑ/-/ɔ/-/ʌ/ and /æ/-/ɛ/ show a greater degree of spectral overlap, resulting in a greater reliance on duration for their separation. If this explanation is valid, it ought to be possible to simulate certain features of our listening test results with a simple pattern classifier. In particular, we were interested in determining whether a simple pattern recognition model would show a greater sensitivity to duration for /ɑ/-/ɔ/-/ʌ/ and /æ/-/ɛ/ than for /i/-/ɪ/, /u/-/ʊ/, and /ɪ/-/e/-/ɛ/.

To test this idea, a quadratic discriminant classifier (Johnson and Winchern, 1982) was trained on measurements from the Hillenbrand *et al.* (1995) database. Excluded from the trained data were (1) tokens with identification error rates of 15% or higher, and (2) tokens with missing values for any of the parameters that were used in discriminant analyses. The parameter set consisted to duration, F_0, and F_1-F_3 sampled at 20% and 80% of vowel duration. [Justification for this particular choice of parameters can be found in Hillenbrand *et al.* (1995) and Hillenbrand and Nearey (1999)]. The pattern recognizer was then tested on measurements from the four different versions of the 300 /hVd/ utterances that were used in the listening tests. The four versions of each utterance were identical with respect to the spectral measurements, but duration was set to (a) the original measured value (OD), (b) 272 ms (ND), (c) 144 ms (SD), or (d) 400 ms (LD).

TABLE III. The most frequent changes in vowel classification resulting from vowel shortening or vowel lengthening by the quadratic discriminant classifier. The percentages in the column to the right reflect the number of shifts in vowel identity (OD→SD or OD→LD) divided by the number of opportunities for such shifts to occur.

Vowel shift	Percentage
Effects of vowel shortening	
/ɔ/→/ɑ/ or /ʌ/	54.2
/æ/→/ɛ/	25.0
/ɑ/→/ʌ/	8.0
Effects of vowel lengthening	
/ʌ/→/ɑ/ or /ɔ/	60.0
/ɛ/→/æ/	33.0

Overall correct classification rates were 98.0% for the OD signals, 97.3% for the ND signals, 88.3% for the SD signals, and 87.0% for the LD signals. Compared to the listening tests results in experiment 2, the OD and ND recognition rates are some 2%–3% higher, while the SD and LD rates are about 3%–4% lower. In common with the listening test results, setting duration to a neutral value produced a very small drop in overall recognition accuracy, while shortening produced a drop in accuracy that was very similar to that produced by lengthening. Of greater interest are the results in Table III, which show the most common shifts in vowel identity produced by the pattern classifier. These analyses were carried out in the same way as those reported for the listening test; i.e., they are based on confusion matrices which focused on instances in which the pattern recognizer classified the OD version of a signal correctly but classified the duration-modified version of the same utterance incorrectly. The results in Table III show a number of important features in common with the listening test results from experiment 2 (Table II). In particular, as with human listeners, the most frequent shifts in vowel classification for the shortened signals consisted of /ɔ/ shifting to /ɑ/ or /ʌ/, and /æ/ shifting to /ɛ/. Further, the most frequent shifts in vowel classification for the lengthened signals consisted of /ʌ/ shifting to /ɑ/ or /ɔ/, and /ɛ/ shifting to /æ/. With the exception of the substantially larger percentage of lengthened /ʌ/ tokens that shifted to /ɑ/ or /ɔ/, the percentages of shifts in Tables II and III are rather similar. Not shown in Table III, but of equal importance, the pattern classifier produced no shifts between /i/ and /ɪ/ and no shifts between /u/ and /ʊ/. Among the /ɪ/-/e/-/ɛ/ cluster, the only duration-dependent shifts that were observed consisted of a modest number (12.0%) of lengthened /ɛ/ tokens shifting to /e/.

Overall, the pattern recognition results support the idea that the role of duration in vowel recognition depends not only on the magnitude and consistency of observed durational differences among vowels but also on the degree to which pairs and groups of vowels are well separated on the basis of spectral cues. Vowels such as /i/-/ɪ/, /u/-/ʊ/, and /ɪ/-/e/-/ɛ/ show consistent durational differences in production but are sufficiently well separated on the basis of spectral features that duration has a rather small influence on perceived vowel quality. On the other hand, vowels such as /ɑ/-/ɔ/-/ʌ/ and /æ/-/ɛ/ show a greater degree of overlap in

their spectral properties and, as a consequence, duration plays a more important role in the recognition of these vowels.

V. GENERAL DISCUSSION

There were many features in common between the findings of experiments 1 and 2. In both experiments, the effect of altering vowel duration was seen to be relatively modest overall since the great preponderance of signals were accurately identified under all four duration conditions. Further, the duration-related effects that were observed in both experiments were quite sensible; that is, vowels with long typical durations tended to shift to adjacent vowels with shorter typical durations when shortened. Conversely, vowels with short typical durations tended to shift to adjacent vowels with longer typical durations when lengthened. The primary difference between the two experiments was that the effect of vowel shortening was considerably greater for the formant synthesized stimuli than the sinusoidally synthesized stimuli.

The greater overall effect of duration in experiment 1 is consistent with the hypothesis that motivated experiment 2. Since the spectral cues to vowel quality are not preserved quite as well in the formant synthesis conditions—resulting in signals whose vowel quality is a bit more ambiguous—altering vowel duration was expected to exert a slightly larger influence on vowel identity in experiment 1. Although an effect in this direction was observed, the results were not quite so simple since the primary difference between the two synthesis methods had to do with the greater influence of vowel shortening for the formant synthesized stimuli relative to the sinusoidally synthesized signals. We believe that this discrepancy between the two experiments is consistent with the HN99 findings on the recognition of natural versus formant synthesized /hVd/ syllables. As mentioned previously, HN99 found that the natural signals were somewhat more intelligible than the formant synthesized versions (96.1% versus 89.8%). However, there was not a simple across-the-board drop in intelligibility resulting from formant synthesis. Vowels showing relatively large decreases in intelligibility as a result of the formant synthesis method included /u/ (shifting primarily to /ʊ/), /æ/ (shifting primarily to /ɛ/), and /ɑ/ (shifting primarily to /ɔ/ or /ʌ/). The explanation for this differential effect of formant coding across vowels is as yet unclear. However, we assume that in experiment 1 of the present study there were a number of OD tokens of /u/, /æ/, and /ɑ/ that were closer in vowel quality than their sinusoidal counterparts to their shorter-duration neighbors, /ʊ/, /ɛ/, and /ʌ/. The result was that a decrease in duration was sufficient to induce a change in vowel identity for many of these formant-synthesized signals. The sinusoidally synthesized signals, on the other hand, preserved the spectral cues of the original vowels more accurately and were less likely to shift to adjacent vowels with shorter typical durations.

The general principle that is illustrated by the differences between experiments 1 and 2 is that the measured importance of a particular acoustic cue to a phonetic dimension depends not only on the precision with which that cue is modeled in the test signals but also on the degree to which other cues to that same dimension are faithfully preserved in the those signals. There are other illustrations of this principle in the acoustic phonetics literature. For example, a series of synthetic speech experiments by Raphael and colleagues (Raphael, 1972, 1981, Raphael *et al.*, 1975) suggested that the duration of a preceding vowel was both a necessary and sufficient cue to the voicing of syllable-final consonants. However, subsequent studies using edited natural speech, which preserved the rich cues to final voicing in the vicinity of articulatory closure and/or release, found that alterations in vowel duration alone were unlikely to induce a change in the voicing of the final consonant (e.g., Wardrip-Fruin, 1982; O'Kane, 1978; Hogan and Rozsypal, 1980; Raphael, 1981; Revoile *et al.*, 1982; Hillenbrand *et al.*, 1984). Results such as these, along with the present findings, serve as a reminder that the conclusions that are drawn from acoustic-phonetic studies can depend on the fine details of the stimulus construction methods.

On a related point, it should be noted that the interpretation of these findings is limited by the relatively simple speech material that was employed. It is well known that there is a large, diverse, and often competing set of influences on vowel duration in connected speech (see Klatt, 1976, for a thorough review). In addition to inherent duration, vowel duration is affected by factors such as overall speaking rate, semantic phenomena such as emphatic stress, grammatical effects such as word- and phrase-final lengthening, variations in lexical stress, and phonetic effects such as the voicing property of preceding or following consonants. Given that the test signals used in this study consisted of citation-form CVC syllables, with identical initial and final consonants, it is nearly certain that listeners would attribute a very large share of the variation in vowel duration to the vowel itself. The situation in connected speech is considerably more complicated, so it is possible that the relatively modest duration effects that were observed here would be even smaller in connected speech. On the other hand, it is also possible that the shorter durations and vowel reduction that characterize connected speech might reduce the spectral contrast among vowels, resulting in a greater role for duration than was observed for the citation-form syllables studied here. Extension of methods such as those used in the present study to connected speech would seem to be a fruitful avenue for further work on this question.

ACKNOWLEDGMENTS

This work was supported by the National Institutes of Health under Grant No. DC-01661 to Western Michigan University. Thanks to Terry Nearey for advice on data analysis. The authors are grateful to Keith Kluender and three anonymous reviewers for insightful comments on an earlier draft.

[1]This is an important difference between our sinusoidal synthesizer and the synthesizer described by McAuley and Quatieri (1986). McAuley and Quatieri go to considerable lengths to preserve the original phase relations among spectral components, while in our method phase is ignored, with the important exception of the steps that are taken to prevent phase discontinuities at the boundaries between frames.

[2]HN99 compared natural and formant-synthesized versions of the 300 /hVd/ signals used in the present study in two experiments with separate listener

groups. In experiment 1 of HN99, the natural signals were 95.4% intelligible versus 88.5% for the formant synthesized signals, used both in HN99 and in experiment 1 of the present study. The corresponding figures for the replication in experiment 2 of HN99 were 96.7% versus 91.0%. Averaged across the two replications, the recognition rate was 96.1% for the natural signals—identical to the recognition rate for the sinusoidal signals used in experiment 2 of the present study—versus 89.8% for the formant synthesized versions.

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

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

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

Chomsky, N., and Halle, M. (**1968**). *The Sound Pattern of English* (Harper and Row, New York).

Crystal, T. H., and House, A. S. (**1988**). ''The duration of American-English vowels: An overview,'' J. Phonetics **16**, 263–284.

Daniloff, R. G., Shriner, T. H., and Zemlin, W. R. (**1968**). ''Intelligibility of vowels altered in duration and frequency,'' J. Acoust. Soc. Am. **44**, 700–707.

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

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

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

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

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.

Hillenbrand, J., Ingrisano, D., Smith, B., and Flege, J. (**1984**). ''Perception of the voiced-voiceless contrast in syllable-final stops,'' J. Acoust. Soc. Am. **76**, 18–26.

Hogan, J., and Rozsypal, A. (**1980**). ''Evaluation of vowel duration as a cue for the voicing distinction in the following word-final consonant,'' J. Acoust. Soc. Am. **67**, 1764–1771.

House, A. S., and Fairbanks, G. (**1953**). ''The influence of consonantal environment upon the secondary acoustical characteristics of vowels,'' J. Acoust. Soc. Am. **25**, 105–113.

Huang, C. B. (**1986**). ''The effect of formant trajectory and spectral shape on the tense/lax distinction in American vowels,'' IEEE ICASSP, 893–896.

Johnson, R. A., and Winchern, D. W. (**1982**). *Applied Multivariate Statistical Analysis* (Prentice-Hall, Englewood Cliffs, NJ).

Klatt, D. H. (**1976**). ''Linguistic uses of segmental duration in English: Acoustic and perceptual evidence,'' J. Acoust. Soc. Am. **59**, 1208–1221.

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.

Laver, J. (**1994**). *Principles of Phonetics* (Cambridge U. P., Cambridge).

McAuley, R. J., and Quatieri, T. F. (**1986**). ''Speech analysis/synthesis based on sinusoidal representation,'' IEEE Trans. Acoust., Speech, Signal Process. **ASSP-22**, 330–338.

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.

O'Kane, D. (**1978**). ''Manner of vowel termination as a perceptual cue to the voicing status of post-vacalic stop consonants,'' J. Phonetics **6**, 311–318.

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

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

Raphael, L. (**1972**). ''Preceding vowel duration as a cue to the perception of the voicing characteristic of word-final consonants in American English,'' J. Acoust. Soc. Am. **51**, 1296–1303.

Raphael, L. (**1981**). ''Durations and contexts as cues to word-final cognate opposition in English,'' Phonetica **38**, 126–147.

Raphael, L., Dorman, M., Freeman, F., and Tobin, C. (**1975**). ''Vowel and nasal duration as cues to voicing in word-final stop consonants: Spectrographic and perceptual studies,'' J. Speech Hear. Res. **18**, 389–400.

Revoile, S., Pickett, J. M., Holden, L. D., and Talkin, D. (**1982**). ''Acoustic cues to final-stop voicing for impaired- and normal-hearing listeners,'' J. Acoust. Soc. Am. **72**, 1145–1154.

Roca, I., and Johnson, W. (**1999**). *A Course in Phonology* (Blackwell, Oxford).

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

Strange, W. (**1989**).

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

van Santen, J. P. H. (**1992**). ''Contextual effects on vowel duration,'' Speech Commun. **11**, 513–546.

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

Wardrip-Fruin, C. (**1982**). ''On the status of temporal cues to phonetic categories: Preceding vowel duration as a cue in final stop consonants,'' J. Acoust. Soc. Am. **71**, 187–195.

Watson, C. I., and Harrington, J. (**1999**). ''Acoustic evidence for dynamic formant trajectories in Australian English Vowels,'' J. Acoust. Soc. Am. **106**, 458–468.

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