Neighboring spectral content influences vowel identification

Lori L. Holt[a)]
Department of Psychology, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, Pennsylvania 15213

Andrew J. Lotto
Department of Psychology & Parmly Hearing Institute, Loyola University Chicago, 6525 North Sheridan Road, Chicago, Illinois 60626

Keith R. Kluender
Department of Psychology, University of Wisconsin-Madison, 1202 West Johnson Street, Madison, Wisconsin 53706

(Received 9 December 1999; accepted for publication 27 April 2000)

Four experiments explored the relative contributions of spectral content and phonetic labeling in effects of context on vowel perception. Two 10-step series of CVC syllables ([bVb] and [dVd]) varying acoustically in *F*2 midpoint frequency and varying perceptually in vowel height from [ʌ] to [ɛ] were synthesized. In a forced-choice identification task, listeners more often labeled vowels as [ʌ] in [dVd] context than in [bVb] context. To examine whether spectral content predicts this effect, nonspeech–speech hybrid series were created by appending 70-ms sine-wave glides following the trajectory of CVC *F*2's to 60-ms members of a steady-state vowel series varying in *F*2 frequency. In addition, a second hybrid series was created by appending constant-frequency sine-wave tones equivalent in frequency to CVC *F*2 onset/offset frequencies. Vowels flanked by frequency-modulated glides or steady-state tones modeling [dVd] were more often labeled as [ʌ] than were the same vowels surrounded by nonspeech modeling [bVb]. These results suggest that spectral content is important in understanding vowel context effects. A final experiment tested whether spectral content can modulate vowel perception when phonetic labeling remains intact. Voiceless consonants, with lower-amplitude more-diffuse spectra, were found to exert less of an influence on vowel perception than do their voiced counterparts. The data are discussed in terms of a general perceptual account of context effects in speech perception. © *2000 Acoustical Society of America*. [S0001-4966(00)01908-1]

PACS numbers: 43.66.Lj, 43.71.An, 43.71.Es [CWT]

INTRODUCTION

One of the earliest and most influential demonstrations of effects of context in speech perception focused on the influence adjacent consonants exert on vowel identification in consonant–vowel–consonant (CVC) syllables. For this study, Lindblom and Studdert-Kennedy (1967) synthesized three vowel series. The first consisted of steady-state series varying perceptually from [ʊ]–[ɪ] via manipulation of *F*2 frequency. The second and third series were comprised of these same vowels embedded in time-varying [wVw] and [jVj] contexts. Critically, vowel formant frequencies at the midpoint of the synthetic stimuli were identical across series. Thus if [wVw] and [jVj] contexts affect vowel perception, then listeners' vowel identification functions for these series should be shifted relative to those of the isolated [ʊ]–[ɪ] series.

Lindblom and Studdert-Kennedy's (1967) observations supported this prediction. Listeners more often identified vowels in [wVw] context as [ɪ]. In [jVj] context, listeners labeled the same vowels more often as [ʊ]. These significant findings have since been supported by Nearey (1989), who extended this perceptual evidence to stop-consonant CVC syllables ([bVb] and [dVd]) with vowel sounds ranging from [ɔ]–[ʌ] and [ʌ]–[ɛ].

By what means might adjacent consonants influence vowel perception? A later study, building upon the results of Lindblom and Studdert-Kennedy (1967), suggests one possibility. Using stimuli very similar to those employed by Lindblom and Studdert-Kennedy, Williams (1986) examined listeners' identification responses across three conditions. Subjects in the first condition identified vowels in series varying perceptually from [wʊw]–[wɪw] and [ʊ]–[ɪ] in an AXB task where they judged whether the random stimulus (*X*) was more similar to the first or third member of the triad (*A* or *B*). These comparison stimuli were always series endpoints ([wʊw]/[wɪw] or [ʊ]/[ɪ]). This condition replicated, in part, the task used by Lindblom and Studdert-Kennedy.[1] In the remaining conditions, Williams capitalized on the observation that nonspeech sine-wave stimuli following the frequency trajectories of speech formants may be perceived as either speech or nonspeech depending upon instructions given to listeners (e.g., Bailey *et al*., 1977; Remez *et al*., 1981; Best *et al*., 1981). Both of the remaining conditions used sine-wave tone complexes varying in frequency trajectory to mimic *F*1, *F*2, and *F*3 formant paths of stimuli from the [wʊw]–[wɪw] and the [ʊ]–[ɪ] series. In the second condition, subjects were informed that these stimuli were speech

[a)]Electronic mail: lholt@andrew.cmu.edu

and were instructed to identify them as [ʊ] or [ɪ]. In the third condition, the same stimuli were described as nonspeech (i.e., as a ''chord'') and a separate set of subjects identified stimuli as having ''low'' or ''high'' pitch.

Replicating Lindblom and Studdert-Kennedy (1967), Williams (1986) observed a shift in vowel identification for speech stimuli; subjects more often labeled vowels as [ɪ] in the context of [wVw]. Interestingly, Williams also found an analogous shift in identification of sine-wave complexes that listeners labeled as speech. Mirroring the results for full-spectrum speech stimuli, time-varying sine-wave triads modeling [wʊw]–[wɪw] formant trajectories were more often labeled [ɪ] than steady-state complexes patterned after [ʊ]–[ɪ]. Importantly, though, there was no evidence of an identification shift among subjects who made pitch judgments of the same sine-wave stimuli. This pattern of results, with identification shifts for stimuli identified as speech, but not for acoustically identical stimuli identified on pitch, suggests that phonetic labeling may be critical in understanding effects of context.

There are several reasons to interpret the conclusions arising from these results with reservation, however. For one, it is important to acknowledge that Williams' (1986) conclusions rely on a null result—namely, the failure to demonstrate a context-dependent shift in pitch identification of sine-wave triads. At the very least, conclusions drawn from null hypotheses need be interpreted with care. However, further evidence also advises caution. In an attempt to replicate Williams' results, Mullennix *et al*. (1988) adopted a stimulus set and an identification task identical to those used by Williams. Whereas Williams reported no context-dependent shift in sine-wave pitch identification, Mullennix *et al*. found that listeners' identification of [wʊw]–[wɪw] tone analogues was shifted toward lower frequencies than was their identification of tone complexes modeled after [ʊ]–[ɪ]. Subjects were significantly more likely to label the pitch of the triad as ''high'' for stimuli mimicking [wʊw]–[wɪw] than for those imitating [ʊ]–[ɪ]. Mullennix *et al*. were therefore able to induce a shift in pitch labeling dependent upon context. Moreover, the observed shift for the sine-wave complexes was in the same direction as that elicited by the speech stimuli after which the tone triads were modeled.

Considering the stimuli and methodology of this experiment were replicas of those used by Williams (1986), these results make it difficult to assess whether phonetic labeling indeed plays a role in the influence of consonant context on vowel perception. Furthermore, the observations of Mullennix *et al*. suggest an alternative hypothesis. Listeners' vowel identification functions shifted regardless of whether they labeled the sine-wave triads phonetically or not. Considering that the sine waves modeled some of the putatively important spectral energy in the speech formants, it may be that the spectral content of adjacent stimuli is the key variable in effects of context on vowel perception.

The present experiments investigate the relative contributions of phonetic labeling and spectral content to context effects in vowel perception. If phonetic labeling is fundamental, then nonspeech contexts that mimic spectral characteristics of consonant context (and are not perceived as speech) should not produce shifts in vowel identification. Alternatively, if spectral content is important, nonspeech stimuli should influence vowel identification in a manner that mimics the influence of the consonants they model. Experiments 1–3 examine these competing hypotheses. Experiment 4 addresses the issue in another way. If spectral content is the important variable in understanding effects of context on vowel perception, then there may be cases in which phonetically labeled consonants' influence on vowel identification can be modulated by manipulations of spectral content. Experiment 4 examines this using voicing as a manipulation of spectral content.

I. EXPERIMENT 1

Pilot study replications of Lindblom and Studdert-Kennedy (1967) and Nearey (1989) demonstrated that native American–English subjects find the stop-consonant series ([bVb]) and [dVd] employed by Nearey to be more compelling instances of CVC syllables than the semi-vowel series ([wVw] and [jVj]) of Lindblom and Studdert-Kennedy.[2] Thus experiment 1 utilizes stop-consonant contexts similar to those of Nearey. However, not all aspects of Nearey's stimuli were modeled in the present studies. In the same pilot experiments, American–English speaking subjects were not adept at labeling the vowel [o] used by Nearey. Even for the best exemplars, subjects labeled stimuli as [o] only approximately 85% of the time. This is likely due to the fact that nondiphthongized [o] (familiar in Western Canadian dialects) is not common to Midwest–American speech. Consequently, experiment 1 introduces vowel stimuli spanning a perceptual range from [ʌ] to [ɛ] via manipulation of $F2$ frequency in a test of effects of [bVb] and [dVd] contexts on vowel identification.

A. Method

1. Subjects

Twenty-eight native English-speaking undergraduates at the University of Wisconsin–Madison participated in return for course credit in Introductory Psychology. All subjects reported normal hearing.

2. Stimuli

Two 10-step series of 200-ms stimuli with constant fundamental frequencies of 120 Hz were synthesized with 12-bit resolution using the cascade branch of the speech synthesizer developed by Klatt (1980). Series varied perceptually from [bʌb]–[bɛb] and [dʌd]–[dɛd]. Synthesis was implemented on a Pentium microcomputer at a sampling rate of 10 kHz. Figure 1 provides schematic spectrograms of endpoint stimuli for each series.

As is illustrated in the figure, formant frequencies are symmetric around the stimulus midpoint (100 ms) and formants $F1-F3$ have a curvilinear trajectory. The trajectory from stimulus onset to midpoint was specified by an equation introduced by Lindblom and Studdert-Kennedy (1967) and used by Nearey (1989),

$$F(t)=F_v+(F_i-F_v)[(t-t_v)^p/t_v^p], \quad (1)$$

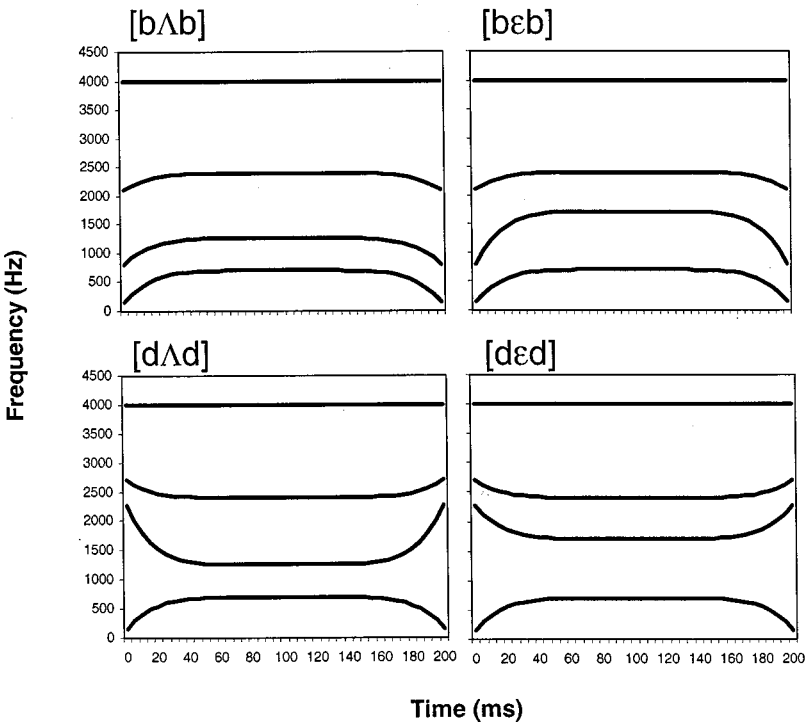

FIG. 1. CVC Series Endpoints. Representative pseudo-spectrograms of experiment 1 stimulus-series endpoints. The top row represents [bVb] stimuli. The bottom row shows [dVd] endpoints. The left column corresponds to low-$F2$ ([Λ]) stimulus endpoints and the right column depicts high-$F2$ ([ɛ]) endpoints.

where $F(t)$ is the frequency of the formant at time t, t_v is the midpoint of the stimulus, F_v is the formant frequency at stimulus midpoint (time t_v), F_i is the formant frequency at stimulus onset, and p is an exponent determining the shape of the formant-frequency trajectory. Here $p=6.0$.[3] The second half of each stimulus was a mirror image of the first.

Onset, offset, and midpoint nominal frequency values for each formant of the endpoint stimuli are displayed in Table I. Within each series, onset and offset frequencies were equal across series members; only $F2$ midpoint frequency varied.

Stimulus presentation and response collection were under the control of an 80486-25 microcomputer. Following D/A conversion (Ariel DSP-16), stimuli were low-pass filtered (4.8 cutoff frequency, Frequency Devices, #677), rms matched in amplitude, amplified (Stewart HDA4), and presented to subjects via headphones (Beyer DT-100) at a level of 70 dB SPL(A).

B. Procedure

Listeners participated in a 2AFC identification task. One to three subjects were tested concurrently in individual sound-attenuated chambers during a single experimental session. In segments mixed across series ([bVb] and [dVd]), participants first identified the vowel as [ʌ] or [ɛ] by pressing either of two buttons on a handheld electronic response box with buttons labeled ''PUTT'' and ''PET'' for reference. Next, listeners identified the consonant as either ''B'' or ''D.'' Although vowel identification was the focus of the experiment, consonant identification responses were elicited to assure that listeners perceived synthetic renditions of CVCs as reasonably good examples of the intended syllables. This also ensured that listeners explicitly phonetically labeled the consonant contexts as well as the vowels.

Participants responded to each stimulus 20 times; order of stimulus presentation was randomized. In all, subjects identified 400 stimuli (10 stim/series×10 repetitions/stim×2 series×2 blocks). The entire session lasted approximately 45 min.[4]

C. Results and discussion

Figure 2 displays mean vowel identification curves for experiment 1.[5] Data were submitted to a 2×10 (consonant context×stimulus-step) within-subjects analysis of variance (ANOVA). As predicted from earlier reports (Lindblom and Studdert-Kennedy, 1967; Nearey, 1989), there was a significant effect of consonant context on vowel identification ($F_{(1,27)}=43.09$, $p<0.0001$). Subjects differentially labeled vowels dependent on consonant context; vowels in [dVd] context were identified as [ʌ] more often than vowels with

TABLE I. Nominal formant frequency values used in synthesis of experiment 1 CVC stimuli.

	Context	Onset	Midpoint	Offset
$F1$	All CVC	150	700	150
$F2$	[bΛb]	800	1260	800
	[bɛb]	800	1710	800
	[dΛd]	2270	1260	2270
	[dɛd]	2270	1710	2270
$F3$	[bVb]	2100	2400	2100
	[dVd]	2700	2400	2700
$F4$	All CVC	4000	4000	4000

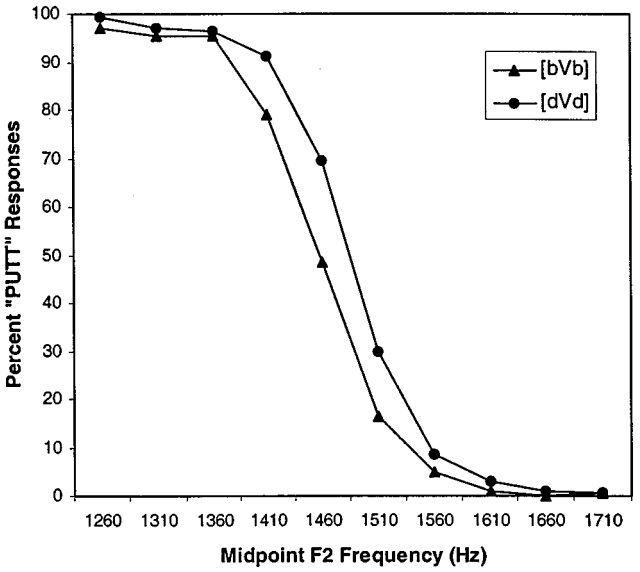

FIG. 2. Identification of Vowels in CVC Context. Mean identification functions for experiment 1. Percent ''PUTT'' responses as a function of midpoint $F2$ frequency are plotted by consonant context.

identical $F2$ midpoint frequencies in [bVb] context. Flanked by high-frequency $F2$ onset–offset consonants ([dVd]), vowels were significantly more likely to be identified as the low $F2$ frequency vowel [ʌ]. In the context of lower $F2$ frequency onset/offset consonants ([bVb]), vowels were more likely to be identified as those with higher $F2$ frequencies (i.e., as [ɛ]). These data thus support findings of earlier studies (Lindblom and Studdert-Kennedy, 1967; Nearey, 1989) and provide a foundation from which to examine the potential influence of nonspeech context upon perceived vowel identity.

II. EXPERIMENT 2

Experiment 2 explores the relative influence of phonetic labeling and spectral content on the effect observed in experiment 1. If spectral content of adjacent context is significant, then any salient acoustic energy, whether speech or nonspeech, should elicit shifts in vowel identification. Experiment 2 stimuli consist of synthetic vowels flanked by nonspeech stimuli that mimic some of the spectral characteristics of the CVC stimuli from experiment 1. In this way, phonetic labeling of context stimuli is eliminated, but some elements of spectral content of CVC context is preserved.

The primary acoustic dimension distinguishing consonant contexts in experiment 1 was $F2$ onset/offset frequency and $F2$ transition trajectories. Experiment 2 models this energy with nonspeech sine-wave glides that mimic the $F2$ frequency trajectories of the CVC stimuli. If spectral content is important, vowels presented in nonspeech contexts with higher-frequency acoustic energy (those modeled after [dVd]) should more often be labeled as [ʌ] than the same vowels flanked by lower-frequency nonspeech contexts modeled after [bVb]. However, nonspeech sounds should not be phonetically labeled. If phonetic labeling is important in effects of context on vowel perception, nonspeech contexts should not shift vowel identification.

A. Method

1. Subjects

Thirty native English-speaking University of Wisconsin undergraduates with normal hearing participated in return for course credit.

2. Stimuli

Experiment 2 utilized a 10-step series of 60-ms steady-state vowels synthesized according to the methods described for experiment 1.[6] As in experiment 1, series members differed only in $F2$ frequency (1260–1710 Hz, in 50-Hz steps) and varied perceptually from [ʌ] to [ɛ]. First, third, and fourth formant frequencies were constant across series members ($F1=700$ Hz, $F3=2400$ Hz, and $F4=4000$ Hz). All formants were steady state over the entire 60-ms duration.

Members of the synthetic vowel series were the basis for two series of nonspeech–speech hybrid stimuli. Each hybrid stimulus was made up of a 70-ms frequency-modulated (FM) sine-wave glide abutted in time to a member of the synthetic vowel series that, in turn, abutted a second 70-ms FM glide that was a mirror image of the first. Frequency trajectories of the sine-wave glides mimicked the $F2$-frequency trajectories of experiment 1 stimuli. For example, the hybrid series modeled after experiment 1 [bVb] series was composed of a nonlinear FM glide from 800 Hz to the $F2$ midpoint frequency (for the vowel) followed by a member of the synthetic vowel series and a final nonlinear FM glide from $F2$-midpoint frequency to 800 Hz. Stimuli mimicking [dVd] series were similar except FM glide onset/offset frequency was 2270 Hz. Equation (1) determined the precise trajectory of each FM glide from midpoint to steady-state values. Glides and vowels were appended online to create a nonspeech–speech hybrid stimulus to model each of [bVb] and [dVd] stimuli of experiment 1.

Pseudo-spectrograms of representative hybrid stimuli are illustrated in Fig. 3. It is important to note that Fig. 3 conveys only the frequency *trajectory* of the stimuli. Although the frequency trajectory of nonspeech precursors is identical to the $F2$ transitions of CVC stimuli from experiment 1, these stimuli should not be considered to be perceptually or acoustically equivalent to the speech that they model. The nonspeech precursors, though modeling putatively important characteristics of CVC context, differ considerably in overall spectral content and are not perceived as speechlike.[7]

The CVC stimuli possess a rich harmonic structure, with energy at each multiple of the fundamental frequency ($f0=120$ Hz). The FM glide, in contrast, possesses energy across only a limited band of frequencies, with no fine harmonic structure and no energy mimicking $F1$ or $F3$. Moreover, formant transitions like those of the CVC stimuli of experiment 1 are not truly frequency-modulated (as are the glide stimuli) because the actual component frequencies do not vary; only the amplitude peak of the spectral envelope varies for formant transitions. Despite these differences, the very simple FM glides capture some of the putatively important spectral energy in the region of CVC $F2$ frequencies. If spectral content accounts for context effects in vowel percep-

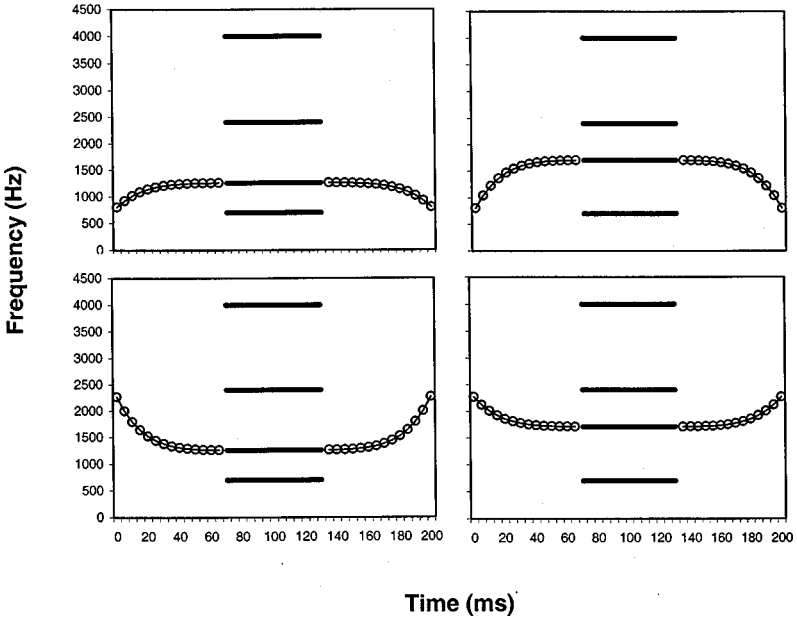

FIG. 3. Glide-Vowel-Glide Series Endpoints. Representative pseudo-spectrograms of experiment 2 stimulus-series endpoints. The format follows that of Fig. 1. The top row represents nonspeech–speech hybrid stimuli modeling the [bVb] series of experiment 1 whereas the bottom row shows endpoints mimicking the [dVd] series. The left column corresponds to low-$F2$ ([ʌ]) stimulus endpoints. The right column depicts high-$F2$ ([ɛ]) endpoints. In each frame, full-formant synthetic speech stimuli are illustrated as solid lines and nonspeech FM glides are shown as stippled lines.

tion, then these very simple nonspeech caricatures, by virtue of their spectral similarity to experiment 1 consonant contexts, may be sufficient to produce effects of context on vowel identification. If phonetic labeling is responsible, then these nonspeech glides should exert no influence on vowel identification.

B. Procedure

Stimulus presentation and response collection were identical to experiment 1. In a 2AFC task, subjects identified the vowel of each stimulus as [ʌ] or [ɛ] by pressing buttons labeled ''PUTT'' or ''PET'' on a handheld electronic response box. After categorizing the vowel, listeners classified the nonspeech segment of the stimulus as ''LOW'' or ''HIGH'' in a task that mirrored the consonant identification task subjects performed in experiment 1. This task also helped to ensure that listeners were not encouraged (perhaps implicitly, by mere participation in a vowel identification task) to attempt to phonetically label the FM glides. Nonspeech–speech hybrid stimuli modeling [bVb] and [dVd] were mixed in stimulus presentation. Each stimulus was presented 20 times in a random order for a total of 400 presentations. Experimental apparatus and procedure were otherwise like that of experiment 1.

C. Results and discussion

Mean identification functions are presented in Fig. 4. A 2×10 (nonspeech context×stimulus-step) ANOVA revealed a significant effect of sine-wave glide context ($F_{(1,29)} = 20.96$, $p < 0.0001$). These results demonstrate that a nonspeech analogue modeling $F2$ frequency trajectory is sufficient to elicit a context effect similar to that obtained with full-formant synthetic speech contexts. In the context of FM glides with higher-frequency onsets and offsets (modeling [dVd]), subjects more often identified vowels as [ʌ], the vowel with a lower $F2$ frequency.

Spectral content thus appears to be an important variable in explaining context effects on vowel perception. Shifts in vowel identification endure even when phonetic labeling is eliminated. Even a very simple sine-wave caricature of a portion of the energy present in rich full-spectrum CVC stimuli is sufficient to influence vowel identification.

One concern that might be raised with respect to these data arises from consideration of auditory grouping principles. Darwin and his colleagues (Darwin, 1984; Darwin and Sutherland, 1984; Darwin *et al.*, 1989), for example, have demonstrated that adjacent nonspeech tones and glides can ''capture'' harmonics of a vowel. Listeners ''attribute'' the energy of the vowel's harmonic to the nonspeech rather than vowel. In their studies, this resulted in an apparent change in $F1$ frequency and a concomitant shift in vowel identification. Considering that Dannenbring (1976) has

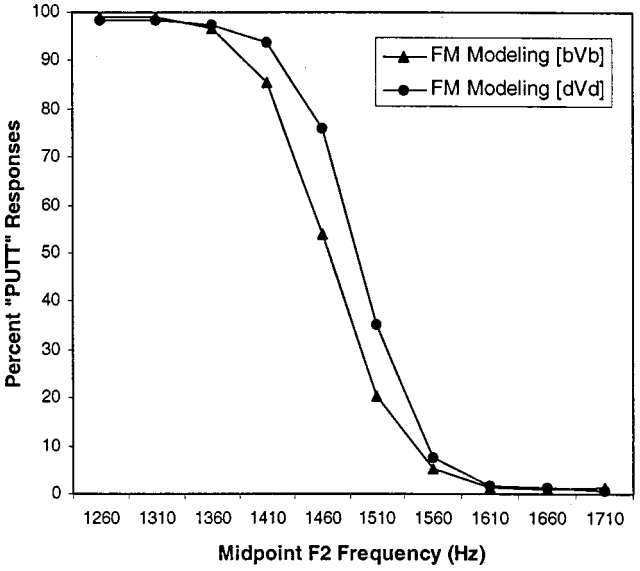

FIG. 4. Identification of Vowels in Glide Context. Mean identification functions for experiment 2. Percent ''PUTT'' responses as a function of midpoint $F2$ frequency are plotted by FM glide context.

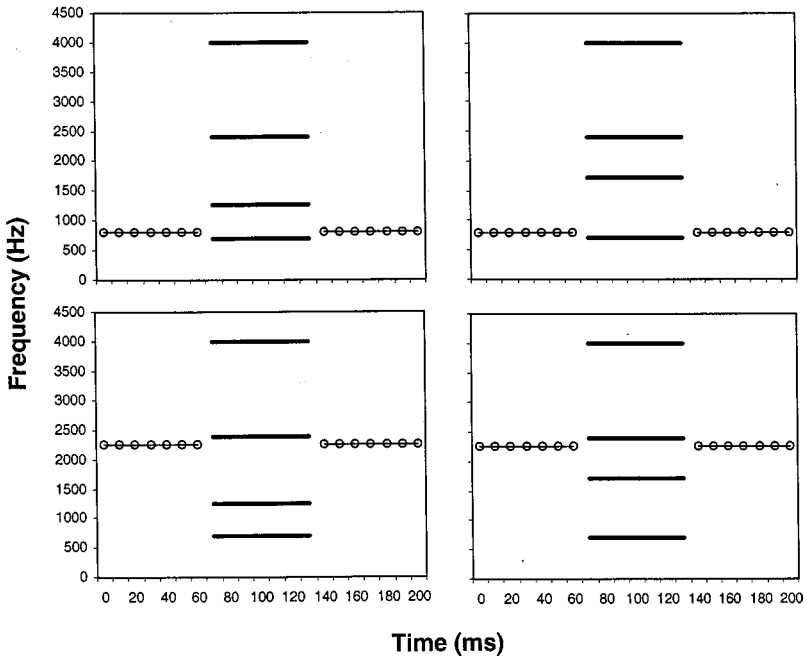

FIG. 5. Tone-Vowel-Tone stimulus series endpoints. Representative pseudo-spectrograms of experiment 3 stimulus-series endpoints. The format follows that of Figs. 1 and 3. The top row represents stimuli that mimic characteristics of [bVb] stimuli; the bottom depicts those that model [dVd] context. The left column illustrates low-$F2$ ([ʌ]) vowels whereas the right column depicts high-$F2$ ([ɛ]) vowels. In each frame, full-formant synthetic speech stimuli are illustrated as solid lines and nonspeech FM glides are shown as stippled lines.

demonstrated that perceived continuity of an intermittent sound is greatest when rising or falling transitions have their discontinuity filled with another sound at the maximum or minimum of their transition trajectory, the stimuli of experiment 2 may have been susceptible to effects of auditory grouping. That is, subjects may have grouped several harmonics around $F2$ with the rising and falling sinusoidal context, thus leading to a perceived shift in $F2$.

III. EXPERIMENT 3

It is worth noting that experiment 2 stimuli were constructed to minimize effects of grouping inasmuch as was possible within this particular stimulus paradigm. Both the nominal vowel $F2$ frequency and the maximum/minimum frequencies of the FM glides that were temporally closest to the vowel were positioned such that they did not fall upon harmonics of the fundamental frequency. To even further reduce potential influences of auditory grouping, experiment 3 used much simpler nonspeech context stimuli. For this new set of nonspeech–speech hybrid stimuli, the nonspeech context was a simple steady-state sine-wave tone situated at the onset/offset frequency of either [bVb] or [dVd] stimuli from experiment 1 (i.e., 800 or 2270 Hz). The tone–vowel–tone stimuli of experiment 3 mitigate grouping of nonspeech energy with harmonics critical to vowel perception because tone context stimuli are closer in frequency to $F1$ and $F3$ (for contexts modeling [bVb] and [dVd], respectively) than $F2$. Thus ''harmonic capture,'' should it be a factor, is much less likely to influence the $F2$ region of the vowel spectra that distinguishes stimuli. Further, because these steady-state tones model even more limited characteristics of CVC spectral characteristics than the FM glides of experiment 2, this stimulus paradigm allows an examination of how closely spectral characteristics of nonspeech analogues must model CVCs to produce shifts in vowel identification.

A. Method

1. Subjects

Thirty-one native English-speaking undergraduate students from the University of Wisconsin–Madison participated in partial fulfillment of the requirements of Introductory Psychology. All listeners reported normal hearing.

2. Stimuli

The 10-step series of steady-state vowel stimuli from experiment 2 were utilized again in experiment 3. For this experiment, steady-state-sine-wave tones, rather than FM glides, served as context stimuli. Each stimulus was comprised of a 70-ms constant-frequency sine-wave tone immediately followed by a member of the 60-ms synthetic vowel series in turn followed by a second 70-ms tone identical to the first. Tones and vowels were appended online to create a model of each of the experiment 1 stimuli. For stimuli modeling [bVb] series members, tone frequency was equivalent to the $F2$ onset/offset frequency of [bVb] stimuli, 800 Hz. Tones with frequencies of 2270 Hz modeled [dVd] series stimuli. Endpoint stimuli are illustrated in Fig. 5.

B. Procedure

In a 2AFC task, listeners identified the vowel of each stimulus as [ʌ] or [ɛ] by pressing buttons labeled ''PUTT'' or ''PET'' on a handheld electronic response box. After categorizing the vowel, subjects classified the nonspeech segment as ''LOW'' or ''HIGH'' in accord with the tasks of experiments 1 and 2. Stimulus presentation was mixed across the two series and each stimulus was presented 20 times in a random order for a sum of 400 total stimuli. Experimental apparatus and procedure were otherwise like that of experiments 1 and 2.

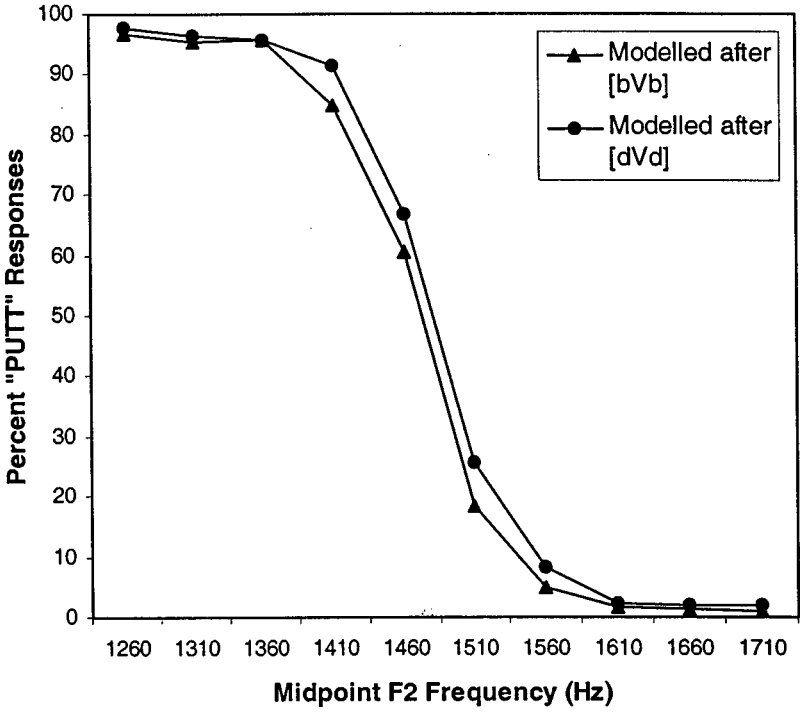

FIG. 6. Identification of Vowels in Tone-Vowel-Tone Context. Mean identification functions for experiment 3. Percent ''PUTT'' responses as a function of midpoint $F2$ frequency are plotted by tone context.

C. Results and discussion

Mean identification functions for experiment 3 are illustrated in Fig. 6. The results of experiment 3 extend the generality of the perceptual effects observed in experiments 1 and 2 to include constant-frequency tones. A static pure tone appended to the onset and offset of a vowel was sufficient to produce a small, but quite reliable, shift in vowel identification ($F_{(1,30)}=8.15$, $p<0.05$). Vowels were more likely to be identified as [ʌ] (having low-frequency $F2$) when flanked by high-frequency tones equivalent to the onset/offset frequency of [dVd] series stimuli in experiment 1.

Although steady-state sine-wave stimuli exert an influence on vowel identification, their effect is smaller than the effects of context observed for consonant and FM glides. This may be because these stimulus hybrids were subject less to auditory grouping, but there are also at least several other explanations for why tones may have relatively less influence on vowel identification. For experiment 3, frequency of the tones was equivalent to onset and offset frequencies of experiment 1 CVCs. Although this seems to be a reasonable choice, it may not have been optimal for producing the largest effects of context. Evidence from psychophysics, for example, suggests listeners perceive brief glides to be a sort of average of frequency across glide trajectory (Brady *et al.*, 1961; Nábělek *et al.*, 1970, 1973). Given this work, it could be argued that tone analogues at mid-trajectory may have more closely modeled context stimuli of experiments 1 and 2. Future psychophysical work that maps the extent of context effects with hybrid speech–nonspeech sounds ultimately will demonstrate whether effect size is sensitive to precise placement of context stimuli and whether these functions can be predicted from psychoacoustics. A second hypothesis is that the dynamic nature of the FM glides from experiment 2 contributes to the influence upon context effects. Given the results of experiment 3, kinematic information offered by dynamic stimuli cannot explain these context effects entirely, but it is possible that this information is a factor. Another difference between the glide and tone stimuli was the overall spectral content. Owing to frequency modulation, glide stimuli possess a richer spectrum than do their tone counterparts. It is conceivable that this spectral distinction influenced the magnitude of the context effect produced by tone stimuli. The results of Shigeno and Fugisaki (1979) suggest that at least one class of context effect becomes more intense as stimulus complexity increases.

IV. EXPERIMENT 4

Time-varying sinusoids and steady-state tones, neither labeled phonetically, were sufficient to shift listeners' identification of vowel stimuli. Taken together, results of experiments 2 and 3 suggest that when phonetic labeling is absent, spectral content is an important variable in understanding how context affects vowel perception. However, it remains at issue whether spectral content is as important when phonetic labeling is intact.

Experiment 4 addresses this question using voicing as a means of manipulating spectral content, while maintaining phonetic labeling. Spectrally, voiced and voiceless consonants vary along many acoustic dimensions. Several distinctions are convenient for the present hypothesis. Namely, voiceless consonants tend to be lower in amplitude than voiced consonants and, owing to the aspiration that often accompanies voicelessness, voiceless consonants tend to have more diffuse spectra than their voiced counterparts. The results of experiments 2 and 3 suggested that acoustic energy in the region of $F2$ frequency trajectories of CVC syllables is important in predicting effects of context. Given this significance, one might predict smaller effects of context from

stimuli with less prominent acoustic energy. By this view, voiceless consonants, with lower-amplitude harmonics and more diffuse spectra, should have relatively less influence on vowel identification than their voiced consonant complements. Experiment 4 examines this hypothesis.

A. Method

1. Subjects

Fourteen undergraduate Introductory Psychology students with normal hearing participated in return for course credit. All listeners learned English as a first language.

2. Stimuli

Two pairs of CVC stimulus series were used in experiment 4. One pair was identical to the [bʌb]–[bɛb] and [dʌd]–[dɛd] series used in experiment 1, except that their duration was 100 ms. This shorter duration was used because previous research (Lindblom and Studdert-Kennedy, 1967) has demonstrated that CVC context effects are smaller for longer stimuli. The present experiment investigates whether a spectral change introduced by manipulating voicing changes the degree to which consonants influence vowel identification. The experimental manipulation is predicted to produce smaller context effects. Therefore, it is desirable to elicit the largest effect possible while maintaining adequate phonetic identification. Shortening the overall stimulus duration to 100 ms thus allowed a more sensitive test of the prediction. In addition, this change allowed experiment 4 to test whether the effect of stimulus duration reported for [jVj] and [wVw] contexts (Lindblom and Studdert-Kennedy, 1967) extends to stop-consonant CVCs.[8]

Duration and formant frequency trajectories for the second pair of series were identical to those of the first. The critical distinction between pairs was that, for the second pair, consonant context was synthesized to mimic voiceless stop consonants rather than voiced consonants as in experiment 1. Thus these stimuli varied perceptually from [pʌp]–[pɛp] and [tʌt]–[tɛt]. It is important to note that the formant frequency trajectories for [bVb] and [pVp] series were identical. Likewise, [dVd] and [tVt] series shared equivalent formant frequency trajectories (as in Fig. 1, except stimuli are 100 ms). Voicing source was the key difference between pairs. Synthesis of this distinction was accomplished by changing the synthesis parameters (Klatt, 1980; parallel branch) of the voiceless series to model the aperiodic aspirated source typical of voiceless consonants. Specifically, the voicing source (Klatt parameter AV) was substituted with an aperiodic source (AH). This change had the effect of lowering the amplitude of the formants during the period of voicelessness. Klatt (1980) synthesizer parameters AV and AH were manipulated carefully so that the transitions of voiceless stimuli were approximately 6 dB lower in amplitude than the same segments of their voiced counterparts. To improve perceptual salience of the voiceless stops, $F1$ cutback (typical of voiceless stop consonants) was also modeled by reducing the amplitude of $F1$ to zero the voiceless portion of the token. No consonant bursts were added to these stimuli. VOT for all stimuli was 40 ms, a value intermediate that of typical natural English [p] and [t] productions (Lisker and Abramson, 1964). Thus voiced and voiceless series shared formant frequency trajectories, but differed in amplitude, periodicity, and low-frequency ($F1$) energy during the voiceless portion of the stimulus.

A final note is in order regarding the voiceless stimulus series used in this experiment. In synthesizing speech for use in controlled laboratory experiments, there is always a compromise between stimulus naturalness and experimental control. In the present case, it was very important to be meticulous with regard to stimulus characteristics because the hypothesis under investigation is coupled closely with the spectral content of the stimuli. With regard to stimulus naturalness, the present investigation demands only that the syllables be perceived phonetically and that listeners' ability to label voiced versus voiceless stimuli is not disproportionate. Thus although the voiceless stimuli employed here do not precisely model natural productions of voiceless CVC syllables (e.g., natural voiceless CVCs tend not to be symmetric), the question under investigation dictated that voiced and voiceless stimuli be as spectrally similar as possible, manipulating only very specific characteristics of the consonant context.

To be certain that listeners do, in fact, perceive these voiceless syllables in the intended manner, experiment 4 incorporates a consonant identification pre-test. In comparing listeners' identification of vowels in the context of voiced and voiceless consonants, it is important to ensure that any observed differences do not stem merely from a difference in the ability of listeners to label voiced versus voiceless consonants; differences in phonetic labeling for voiced versus voiceless consonants may influence the degree to which spectral content can be held responsible for effects of context on vowel identification.

B. Procedure

Therefore, each listener first participated in a preliminary task in which consonant identification was assessed. During this task, listeners heard each stimulus from each 10-member stimulus series ([bVb], [dVd], [pVp], [tVt]) five times for a total of 200 stimuli. On each trial, subjects labeled the consonants of the syllables as ''B,'' ''D,'' ''P,'' or ''T.''

Following this task, each listener completed two blocks of trials to assess vowel identification in the context of adjacent consonants. The trials were blocked by voicing, such that listeners identified vowels in the context of [bVb] and [dVd] separate from vowels in [pVp] and [tVt] contexts. The order of blocks was counterbalanced across subjects. Within a block, subjects heard 20 repetitions of each member of each stimulus series. On every trial, subjects indicated perceived vowel identity as [ʌ] or [ɛ] by pressing one of two buttons labeled ''HUD'' and ''HEAD,'' then responded to the consonant as ''P/T'' or ''B/D.''[9]

C. Results and discussion

Subjects performed very well on the consonant identification pretest. On average, subjects correctly labeled conso-

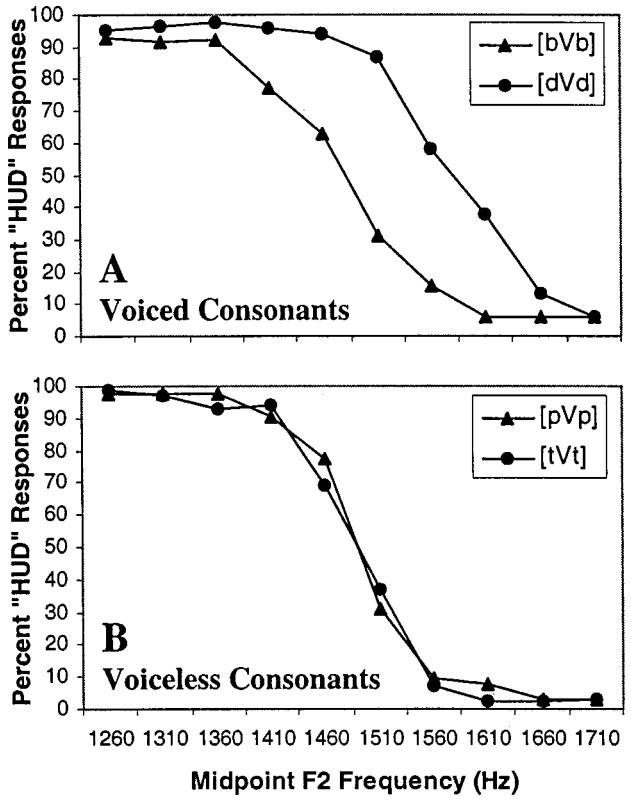

FIG. 7. Identification of Vowels in ±Voice CVC Contexts. Panel A illustrates listeners' identification of vowels in the context of voiced consonants. Panel B depicts the same listeners' identification responses for vowels in the context of voiceless consonants.

nants from [bVb], [dVd], [pVp], and [tVt] series 95.3%, 91.2%, 93.3%, and 90.7% of the time, respectively. Thus despite strict constraints on stimulus construction, all consonant contexts were readily identifiable.

The influence of voiced consonants on vowel identification is shown in the top panel of Fig. 7. As in experiment 1, voiced consonant context elicited a substantial effect on vowel identification ($F_{(1,13)}=73.18$, $p<0.0001$). In the context of [dVd], vowels were significantly more likely to be identified as [Λ]. Consistent with earlier observations (Lindblom and Studdert-Kennedy, 1967), the magnitude of the context effect produced by these shorter stimuli (100 ms versus 200 ms) was greater than that observed for experiment 1, thus extending effects of duration upon phonetic context effects to stop-consonant CVCs.

The significant question of experiment 4 was whether voiceless consonant contexts, though easily labeled and possessing the same formant trajectories as their voiced counterparts, would exhibit relatively less influence on vowel identification due to the differences in spectral content between series. The lower panel of Fig. 7 depicts the data pertinent to this question. As is clear from the graph, voiceless consonants were substantially less effective in shifting vowel identification than their voiced counterparts. In fact, there is no significant influence of neighboring voiceless consonants upon vowel identification at all ($F_{(1,13)}=0.66$, $p=0.435$). These data thus strongly implicate spectral content as an important factor in context effects in vowel perception, even in stimulus paradigms were phonetic labeling is intact.

These results are especially interesting because it is unlikely that extant accounts of context effects would forecast these outcomes. Most theoretical accounts of context effects in speech perception, whether embodied by modular processes (Liberman *et al.*, 1957; Liberman and Mattingly, 1985; Mann, 1980, 1986), Direct Realist accentuation of distal events (Fowler *et al.*, 1990; Fowler, 1986, 1996), or reference to tacit knowledge (Repp, 1982, 1983) suggest that mechanisms of context effects in speech perception are intimately linked with speech production generally, and to effects of coarticulation more specifically. However, the dynamics of articulation for voiced/voiceless consonants at the same place of articulation should be very similar because the voicing distinction is related primarily to a change in the timing relationship between release of oral closure and onset of vocal fold vibration, not movement of supra-laryngeal articulators (Lisker and Abramson, 1964, 1971). Thus articulatory patterns for voiced versus voiceless stimuli should be quite similar and their coarticulatory influences on neighboring sounds should be analogous, too. Given this similarity, it is unclear how accounts that rely on knowledge of or recovery of coarticulatory aspects of speech production to aid perception would account for the results of experiment 4.

V. GENERAL DISCUSSION

The goal of the present work was to assess the relative contributions of phonetic labeling and spectral content on context effects in vowel categorization. Williams (1986) reported data suggesting that effects of context on vowel identification are dependent upon phonetic identification. Listeners exhibited context effects for sine-wave speech in conditions where sine-wave triads were phonetically labeled, but not in conditions where the same sounds were labeled on pitch. Mullennix *et al.* (1988) later called the necessity of phonetic labeling into question by reporting a context effect for sine wave triads labeled either phonetically or with pitch labels, thus suggesting a possible influence of spectral content.

The present data support this possibility. Spectral content, rather than phonetic labeling, seems to be the chief characteristic contributing to context effects in vowel perception. The results of experiments 2 and 3 indicate that nonspeech acoustic contexts with spectral characteristics modeled after consonant contexts of experiment 1 are sufficient to produce shifts in vowel identification. Frequency-modulated sine-wave glides and, to a lesser extent, steady-state sine-wave tones, elicit shifts in listeners' vowel identification responses. Moreover, these shifts are in accord with patterns of perception produced by consonant neighbors. These nonspeech contexts were not phonetically labeled, so it appears that the limited spectral characteristics of the consonant acoustics that they modeled are adequate to predict effects of context on vowel identification. Furthermore, because these nonspeech stimuli elicited context effects, the data suggest that close correspondence to productions of the human vocal tract is not a prerequisite for context effects in speech perception.

Experiment 4 linked results of experiments 2 and 3 back to more ordinary speech perception in CVC stimuli. Voiced,

but not voiceless, consonant context modulated listeners' vowel identity responses. This is a curious finding in that extant theoretical accounts of speech perception that rely upon links to speech production to account for context effects would not predict such an interaction. Several new empirical investigations, however, suggest a manner by which to interpret these results. Recently, there have been a number of reports investigating explicitly the possibility that the influence of neighboring context on speech identification may be a consequence of general auditory processes (Holt, 1999; Lotto *et al*., 1997; Lotto and Kluender, 1998; for an overview see Holt and Kluender, in press). These studies have provided two primary lines of evidence to suggest effects of adjacent neighbors on speech perception may have origins in very general auditory mechanisms.

The first line of evidence ties closely with results of experiments 2 and 3 and involves correspondence between the effects of neighboring speech and nonspeech sounds upon perception of speech. Lotto and Kluender (1998) demonstrated that sine-wave FM glides modeling $F3$ transitions of [al] or [ar] and steady-state tones situated at the $F3$ offset frequencies of [al] or [ar] induce the same pattern of [ga]–[da] identification responses as do natural and synthetic speech tokens of [al] and [ar]. Likewise, Holt (1999) has reported that listeners exhibit shifts in identification of CV syllables when the CVs are preceded by [i], [u], or by nonspeech precursors modeling $F2$ characteristics of these vowels. The correspondence of speech and nonspeech–speech hybrid context effects suggests that general auditory alternatives to speech-specific explanations may be tenable and that these mechanisms are related to spectral content of the acoustic signal.

A second line of evidence bolsters this account. Lotto *et al*. (1997) trained Japanese quail (*Coturnix coturnix japonica*) to peck to a lighted key in response to endpoints of the same [da]–[ga] series for which humans had exhibited context-dependent shifts in identification (Lotto and Kluender, 1998). Birds trained to peak to [ga] pecked most vigorously to novel intermediate members of the [da]–[ga] series that were preceded by [al]. Correspondingly, [da] positive quail pecked most robustly when novel stimuli were preceded by [ar]. Japanese quail thus exhibited a context effect analogous to human shifts in CV identification. Quail are unlikely to have relied upon recovery or representation of human vocal tract dynamics. Furthermore, the quail had no experience with coarticulated speech, so their behavior cannot be explained based on learned covariance of the acoustic attributes of coarticulated speech.

These twin lines of evidence favor a general auditory account of context effects in speech perception rather than explanations hinging on correspondences between speech perception and speech production or speech-specific mechanisms. Taken together, these data suggest spectral content and its influence upon the auditory system may be critical variables in explaining context effects. Thus these data join previous reports of speech perception phenomena that appear to arise from general perceptual processes and are not dependent upon phonetic labeling (e.g., Diehl and Walsh, 1989; Lotto *et al*., 1996; Sinnott *et al*., 1998).

Discussion of previous results (Holt, 1999; Lotto and Kluender, 1998; Lotto *et al*., 1997) has cast these findings as cases of spectral contrast. Considered in this way, the results of experiments 1–3 could be interpreted in the following manner: In the context of higher-frequency spectral energy (i.e., [dVd] stimuli and the nonspeech analogues that model them), vowels are perceived as the alternative with lower-frequency spectra (i.e., as [Λ]). This pattern of perception is typical of phonetic context effects in speech and has been noted by other authors as well (e.g., Lindblom and Studdert-Kennedy, 1967; Repp, 1983; Fowler *et al*., 1990). However, the present results implicate general auditory mechanisms in explanations of context effects and, therefore, suggest that contrast may serve as more than a mere description. Mechanisms of contrast that enhance change in the acoustic signal potentially could produce perceptual results like those reported here.

This hypothesis is all the more feasible considering that contrast is an important mechanism for exaggerating differences between neighboring objects and events across perceptual modalities. The best-known examples are in the visual domain (enhancement of edges produced by lateral inhibition, Hartline and Ratliff, 1957; lightness judgments, Koffka, 1935; judgment of line orientation, Gibson, 1933), but context effects in behavior are observed for all sensory modalities (von Bekesy, 1967; Warren, 1985). Across domains, contrast is a central characteristic of mechanisms that serve to exaggerate change in the physical stimulus and to maintain an optimal dynamic range. Perceptual contrast, in this case spectral contrast, may play an important role in perception of speech, too.

This perspective also contributes an explanation for the curious results of experiment 4. Indulging the hypothesis that general perceptual mechanisms of spectral contrast may play a role in vowel context effects, it is useful to consider how the auditory system might implement spectral contrast. Delgutte (1996; Delgutte *et al*., 1996), for example, has suggested spectral contrast might be realized by the auditory system via neural adaptation. Adaptation can be described very generally as a decrease in neurons' discharge rate following an initial response, such that responses to subsequent stimuli are depressed (Harris and Dallos, 1979; Smith, 1979). Though neural adaptation is perhaps best known for its influence in the auditory system at the level of the auditory (VIIth) nerve, adaptation exists at every level of the auditory system and occurs on many time scales, from a few milliseconds to several seconds or even minutes (Kiang *et al*., 1965; Smith, 1979).

Neurons in the auditory system have a preferred, or characteristic, frequency to which they are most likely respond with an action potential (Brugge and Reale, 1985; Irvine, 1992; Ruggero, 1992). Adaptation may produce contrast because neurons excited by stimulus components spectrally close to their preferred frequency fire and subsequently become adapted. Consider what happens when another stimulus follows. Neurons most sensitive to frequencies present in the preceding stimulus will be likely to have fired in response to it, leaving them relatively less responsive due to adaptation. However, frequencies absent in the pre-

ceding stimulus will tend to be encoded by more responsive *unadapted* neurons. Thus on a population level, there is a shift in neural response toward frequencies absent in the preceding stimulus.

If we entertain neural adaptation as a candidate mechanism for spectral contrast, a prediction arises. All other things being equal, less spectrally distinct or less intense precursors should tend to result in fewer adapted neurons. Thus when a subsequent stimulus follows, less adaptation should result in less of a population shift in neural response. That is, there should be less neural contrast. If this change influences perception, such stimuli should exert a smaller effect of context on their neighbors. For the differences in spectra between voiced and voiceless consonants in experiment 4, the account outlined above predicts that lower-amplitude, more diffuse formant energy of voiceless consonants such as [p] and [t] ought to exert relatively less influence on neighboring vowels than should their voiced counterparts, [b] and [d]. In previous discussions (Lotto *et al*., 1997; Lotto and Kluender, 1998; Holt and Kluender, in press), spectral contrast has served as a useful descriptive convenience in discussing general auditory mechanisms by which context effects might occur. Experiment 4 provides the first perceptual evidence to advise that spectral contrast may provide more than descriptive convenience in explaining context effects in speech perception. Rather, it serves as a framework from which to make novel predictions about perception. Much work remains to be done before the precise means by which context effects in speech perception arise is known. Nonetheless, a number of things are clear, even at this point.

Often, arguments that context effects in speech perception arise from ''general auditory'' processes are taken as suggesting that the mechanisms must be *peripheral*. If, by peripheral, authors mean mechanisms that exert their influence in the cochlea or at the level of the auditory nerve, then this is almost certainly false. Evidence in other stimulus paradigms (e.g., Holt, 1999; Lotto, 1996) suggests that more central regions of the auditory system are likely involved in context effects. In addition, neurophysiological investigation of these effects has borne little evidence of robust neural encoding of context effects at the level of the auditory nerve (Holt and Rhode, 2000). This need not rule out ''general auditory'' accounts. Consider, again, adaptation as a candidate mechanism. As Delgutte (1996) has pointed out, adaptation occurs at all levels of the auditory system and possesses a host of time courses.

It is very important to note that the results of experiment 4 do not demonstrate that neural adaptation serves as a mechanism for context effects in speech perception. As Summerfield and others have argued within other theoretical domains (e.g., Summerfield *et al*., 1984, 1987), effects similar to those putatively accomplished by neural adaptation may be accomplished by adaptation of suppression (a mechanism whereby inhibitory inputs are suppressed, thus enhancing the response of some population of neurons by virtue of decreasing inhibition) or other temporal interaction mechanisms such as disinhibition, long-lasting inhibition, or facilitation. Temporal interactions abound within the auditory system (Delgutte, 1996). Neural adaptation, even if it does play a role, is unlikely to account entirely for such effects. Especially at higher levels of the auditory system, mechanisms of temporal interaction are likely to be more complex than simple adaptation. Even at the cochlear nucleus, responses of single neurons to brief vowel stimuli presented in rapid succession are suppressed when a particular vowel is preceded by another vowel that, by itself, produces no response (Caspary *et al*., 1977). These data suggest that there exist mechanisms distinct from adaptation that are involved in temporal interactions among neighboring speech sounds. The present results, coupled with these neurophysiological clues, suggest that although neural adaptation may not provide a complete account of effects of context in speech perception, spectral contrast has a useful role in predicting perceptual results.

In explaining their original finding that CVC context influences listeners' identification of vowels, Lindblom and Studdert-Kennedy (1967) offered several possible theoretical interpretations. They noted that their data were agreeable to articulation-based theoretical perspectives like Motor Theory (Liberman *et al*., 1957) and Analysis-by-Synthesis (Stevens and House, 1963; Stevens and Halle, 1967). However, Lindblom and Studdert-Kennedy also examined the possibility that more general perceptual mechanisms may be responsible. Indeed, they devoted most of their discussion to potential explanations in terms of general auditory processes. In most all examinations of phonetic context effects since, authors have interpreted their data more categorically as evidence that context effects in speech perception originate in properties of speech perception distinct from its auditory characteristics (e.g., Mann, 1980; Repp, 1982; Williams, 1986; Fowler *et al*., 1990). The present data, in combination with other recent results (Lotto *et al*., 1997; Lotto and Kluender, 1998; Holt, 1999) suggest that general perceptual mechanisms sensitive to similarities in spectral content play an important role in context effects in speech perception.

ACKNOWLEDGMENTS

This work was supported in part by a National Science Foundation Predoctoral Fellowship to the first author. Additional support was provided by NSF Young investigator Award DBS-9258482 to the third author. Some of the data were presented at the 131st Meeting of the Acoustical Society of America in Indianapolis, IN. The authors thank Bjorn Lindblom and Terry Nearey for their helpful comments as this work progressed. Correspondence and requests for reprints should be addressed to Lori L. Holt, Department of Psychology, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213 (email: lholt@andrew.cmu.edu).

[1]One distinction between these studies is that Williams (1986) did not include a [jVj] series. In addition, Lindblom and Studdert-Kennedy employed a two-alternative forced-choice (2AFC) identification procedure. Williams also collected 2AFC data for two of his three conditions. Results were comparable across methods, but the overall effects observed for 2AFC identification were smaller. The ABX results are described here because they provide the most complete account of Williams' findings.

[2]In these pilot studies, we replicated Lindblom and Studdert-Kennedy (1967) and Nearey (1989). However, we were interested in the putative role of phonetic identity in phonetic context effects, so we assessed listeners' ability to identify consonants as well as vowels in the CVC syllables. Though most subjects converged upon consistent labeling of [wVw] and

[jVj] stimuli, some listeners reversed their identifications. The same listeners had little trouble identifying stop-consonant CVCs. Therefore, we chose to use stop-consonant CVCs for the remainder of the study.

[3]Lindblom and Studdert-Kennedy (1967) created CVCs perceptually corresponding to [wVw] and [jVj] using $p=2.0$, a parabolic function. Like Nearey (1989), the present experiment employed $p=6.0$ to produce percepts of CVCs conforming to [bVb] and [dVd]. Acoustically, use of a larger exponent creates formant transitions that traverse frequency more quickly, creating a longer duration at relative steady-state midpoint frequency relative to smaller exponents.

[4]In studies of context effects in speech perception, the method of comparing identification of isolated exemplars to identification of the same exemplars in the context of adjacent stimuli has been used extensively (e.g., Lindblom and Studdert-Kermedy, 1967; Mann, 1980; Mann and Repp, 1980). Despite this common practice, this means may not be the most appropriate manner of assessing ''baseline'' perception in the case of vowel identification. This is to say that, considered in terms of adjacent spectral energy and its influence on perception of vowel stimuli, silence may not be the best control ''stimulus'' for comparison with prominent low- or high-frequency energy. For this reason, context effects are assessed here in terms of the extent to which identification of a vowel stimulus is shifted relative to adjacent context rather than the extent to which it is shifted from listeners' identification of the vowel in isolation.

[5]Listeners were extremely accurate in labeling consonant contexts as ''B'' or ''D'' (96.7% and 97.2% accuracy, respectively), indicating that CVC syllables were perceived as reasonable exemplars of the intended speech sounds. To remain in accord with this first experiment, the remaining studies (experiments 2–4) use similar methods for labeling context. In all cases, listeners performed very well on the labeling tasks. Listeners' responses will be reported in detail only when they are central to the hypotheses of interest.

[6]The value 60 ms was chosen as a compromise between preserving spectral information inherent to the formant transition of CVC syllables and preserving the spectral information of the relatively steady-state vowel portion of the syllables. Inspection of the Klatt synthesis parameters derived from Eq. (1) (updated every 5 ms) led us to choose 60 ms as a compromise between these competing demands. It was also desirable to have the overall stimulus duration of experiment 2 stimuli be equivalent to that of the CVC stimuli of experiment 1 (200 ms). Each nonspeech context stimulus was thus 70 ms, for a total stimulus duration of 200 ms ($70+60+70$ ms $=200$ ms).

[7]Independent of experiment 2, a small group of naïve listeners described these stimuli. No listener described the nonspeech context stimuli as speechlike. The most common responses were ''computer game sounds,'' ''beeps,'' ''electronic music,'' and ''chirps.''

[8]Nearey (1989) used 100-ms stop-consonant CVCs in an analysis of the effect of consonant context on vowel identification. However, this study did not include a direct comparison across syllable duration.

[9]Different response labels were used in this experiment to avoid confounding the labels with the voiceless series.

Bailey, P. J., Summerfield, A. Q., and Dorman, M. (**1977**). ''On the identification of sine wave analogues of certain speech sounds,'' Haskins Lab. Status Report on Speech Res., **SR-51/52**, 1–25.

von Bekesy, G. (**1967**). *Sensory Inhibition* (Princeton University Press, Princeton, NJ).

Best, C. T., Morrongiello, B., and Robson, R. (**1981**). ''Perceptual equivalence of acoustic cues in speech and nonspeech perception,'' Percept. Psychophys. **29**, 191–211.

Brady, P. T., House, A. S., and Stevens, K. N. (**1961**). ''Perception of sounds characterized by a rapidly changing resonant frequency,'' J. Acoust. Soc. Am. **33**, 1357–1362.

Brugge, J. F., and Reale, R. A. (**1985**). ''Auditory cortex,'' in *Cerebral Cortex*, Vol. 4, edited by A. Peters and E. G. Jones (Plenum, New York), pp. 229–271.

Caspary, D. M., Rupert, A. L., and Moushegian, G. (**1977**). ''Neuronal coding of vowel sounds in the cochlear nuclei,'' Exp. Neurol. **54**, 414–431.

Dannenbring, G. L. (**1976**). ''Perceived auditory continuity with alternately rising and falling frequency transitions,'' Can. J. Psychol. **30**, 99–114.

Darwin, C. J. (**1984**). ''Perceiving vowels in the presence of another sound: Constraints on formant perception,'' J. Acoust. Soc. Am. **76**, 1636–1647.

Darwin, C. J., Pattison, H., and Gardner, R. B. (**1989**). ''Vowel quality changes produced by surrounding tone sequences,'' Percept. Psychophys. **45**, 333–342.

Darwin, C. J., and Sutherland, N. S. (**1984**). ''Grouping frequency components of vowels: When is a harmonic not a harmonic?'' Q. J. Exp. Psychol. **36A**, 193–208.

Delgutte, B. (**1996**). ''Auditory neural processing of speech,'' in *The Handbook of Phonetic Sciences*, edited by W. J. Hardcastle and J. Laver (Blackwell, Oxford), pp. 507–538.

Delgutte, B., Hammond, B. M., Kalluri, S., Litvak, L. M., and Carian, P. A. (**1996**). ''Neural encoding of temporal envelope and temporal interactions in speech,'' in *Proceedings of Auditory Basis of Speech Perception*, edited by W. Ainsworth and S. Greenberg, European Speech Communication Association.

Diehl, R. L., and Walsh, M. A. (**1989**). ''Effects of syllable duration on stop-glide identification in syllable-initial and syllable-final position by humans and monkeys,'' J. Acoust. Soc. Am. **85**, 2154–2164.

Fowler, C. A. (**1986**). ''An event approach to the study of speech perception from a direct-realist perspective,'' J. Phonetics **14**, 3–28.

Fowler, C. A. (**1996**). ''Listeners do hear sounds, not tongues,'' J. Acoust. Soc. Am. **99**, 1730–1741.

Fowler, C. A., Best, C. T., and McRoberts, G. W. (**1990**). ''Young infants' perception of liquid coarticulatory influences on following stop consonants,'' Percept. Psychophys. **48**, 559–570.

Gibson, J. J. (**1933**). ''Adaptation, after-effect and contrast in the perception of curved lines,'' J. Exp. Psychol. **16**, 1–31.

Harris, D. M., and Dallos, P. (**1979**). ''Forward masking of auditory-nerve fiber responses,'' J. Neurophysiol. **42**, 1083–1107.

Hartline, H. K., and Ratliff, F. (**1957**). ''Inhibitory interaction of receptor units in the eye of *Limulus*,'' J. Gen. Physiol. **40**, 1357–1376.

Holt, L. L. (**1999**). ''Auditory constraints on speech perception: An examination of spectral contrast,'' unpublished doctoral dissertation, University of Wisconsin–Madison.

Holt, L. L., and Kluender, K. R. (in press). ''General auditory processes contribute to perceptual accommodation of coarticulation,'' Phonetica.

Holt, L. L., and Rhode, W. R. (**2000**). ''Examining context-dependent speech perception in the chinchilla cochlear nucleus,'' Midwinter Meeting of the Association for Research in Otolaryngology.

Irvine, D. R. F. (**1992**). ''Physiology of the auditory brainstem,'' in *The Mammalian Auditory Pathway: Neurophysiology*, edited by A. N. Popper and R. R. Fay (Springer Verlag, New York), pp. 153–231.

Kiang, N. Y. S., Watanabe, T., Thomas, E. C., and Clark, L. F. (**1965**). ''Discharge patterns of single fibers in the cat's auditory nerve,'' Res. Monogr., 35 (MIT Press, Cambridge, MA).

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

Koffka, K. (**1935**). *Principles of Gestalt Psychology* (Harcourt Brace, New York).

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. (**1957**). ''Perception of the speech code,'' Psychol. Rev. **74**, 431–461.

Liberman, A. M., and Mattingly, I. G. (**1985**). ''The motor theory of speech perception revised,'' Cognition **21**, 1–36.

Lindblom, B., and Studdert-Kennedy, M. (**1967**). ''On the role of formant transitions in vowel recognition,'' J. Acoust. Soc. Am. **42**, 830–843.

Lisker, L., and Abramson, A. S. (**1964**). ''A cross-language study of voicing in initial stops: Acoustical measurements,'' Word **20**, 384–422.

Lisker, L., and Abramson, A. S. (**1971**). ''Distinctive features and laryngeal control,'' Language **47**, 767–785.

Lotto, A. J. (**1996**). ''General auditory constraints in speech perception,'' unpublished doctoral dissertation, University of Wisconsin–Madison.

Lotto, A. J., and Kluender, K. R. (**1998**). ''General contrast effects of speech perception: Effect of preceding liquid on stop consonant identification,'' Percept. Psychophys. **60**, 602–619.

Lotto, A. J., Kluender, K. R., and Green, K. P. (**1996**). ''Spectral discontinuities and the vowel length effect,'' Percept. Psychophys. **58**, 1005–1014.

Lotto, A. J., Kluender, K. R., and Holt, L. L. (**1997**). ''Perceptual compensation for coarticulation by Japanese quail (*Coturnix coturnix japonica*),'' J. Acoust. Soc. Am. **102**, 1134–1140.

Mann, V. A. (**1980**). ''Influence of preceding liquid on stop-consonant perception,'' Percept. Psychophys. **28**, 407–412.

Mann, V. A. (**1986**). ''Distinguishing universal and language-dependent levels of speech perception: Evidence from Japanese listeners' perception of English ''l'' and ''r'','' Cognition **24**, 169–196.

Mann, V. A., and Repp, B. H. (**1980**). ''Influence of vocalic context on perception of the [sh]–[s] distinction,'' Percept. Psychophys. **28**, 213–228.

Mullennix, J. W., Pisoni, D. B., and Goldinger, S. D. (**1988**). ''Some effects of time-varying context on the perception of speech and nonspeech sounds,'' Res. on Speech Percept.: Prog. Report No. 14, Indiana University.

Nábělek, I. V., Nábělek, A. K., and Hirsh, I. J. (**1970**). ''Pitch of tone bursts of changing frequency,'' J. Acoust. Soc. Am. **48**, 536–553.

Nábělek, I. V., Nábělek, A. K., and Hirsh, I. J. (**1973**). ''Pitch of sound bursts with continuous or discontinuous change of frequency,'' J. Acoust. Soc. Am. **53**, 1305–1312.

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

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

Repp, B. H. (**1982**). ''Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception,'' Psychol. Bull. **92**, 81–110.

Repp, B. H. (**1983**). ''Bidirectional contrast effects in the perception of VC–CV sequences,'' Percept. Psychophys. **33**, 147–155.

Ruggero, M. A. (**1992**). ''Physiology and coding of sound in the auditory nerve,'' in *The Mammalian Auditory Pathway: Neurophysiology*, edited by A. N. Popper and R. R. Fay (Springer Verlag, New York), pp. 34–93.

Shigeno, S., and Fugisaki, H. (**1979**). ''Effect of a preceding anchor upon the categorical judgment of speech and nonspeech stimuli,'' Japanese Psych. Res. **21**, 165–173.

Sinnott, J. M., Brown, C. H., and Borneman, M. A. (**1998**). ''Effects of syllable duration on stop-glide identification in syllable-initial and syllable-final position by humans and monkeys,'' Percept. Psychophys. **60**, 1032–1043.

Smith, R. L. (**1979**). ''Adaptation, saturation, and physiological masking in single auditory-nerve fibers,'' J. Acoust. Soc. Am. **65**, 166–178.

Stevens, K. N., and Halle, M. (**1967**). ''Remarks on analysis by synthesis and distinctive features,'' in *Models for the Perception of Speech and Visual Form*, edited by W. Wathem-Dunn (MIT Press, Cambridge, MA).

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

Summerfield, Q., Haggard, M. P., Foster, J., and Gray, S. (**1984**). ''Perceiving vowels from uniform spectra: Phonetic exploration of an auditory aftereffect,'' Percept. Psychophys. **35**, 203–213.

Summerfield, Q., Sidwell, A., and Nelson, T. (**1987**). ''Auditory enhancement of changes in spectral amplitude,'' J. Acoust. Soc. Am. **81**, 700–707.

Warren, R. M. (**1985**). ''Criterion shift rule and perceptual homeostasis,'' Psychol. Rev. **92**, 574–584.

Williams, D. R. (**1986**). ''Role of dynamic information in the perception of coarticulated vowels,'' unpublished doctoral dissertation, University of Connecticut.