Effects of later-occurring nonlinguistic sounds on speech categorization

Travis Wade[a] and Lori L. Holt
Department of Psychology and the Center for the Neural Basis of Cognition, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213

(Received 26 December 2004; revised 25 May 2005; accepted 2 June 2005)

Nonspeech stimuli influence phonetic categorization, but effects observed so far have been limited to precursors' influence on perception of following speech. However, both preceding and following speech affect phonetic categorization. This asymmetry raises questions about whether general auditory processes play a role in context-dependent speech perception. This study tested whether the asymmetry stems from methodological issues or genuine mechanistic limitations. To determine whether and how backward effects of nonspeech context on speech may occur, one experiment examined perception of CVC words with [ga]-[da] series onsets followed by one of two possible embedded tones and one of two possible final consonants. When the tone was separated from the target onset by 100 ms, contrastive effects of tone frequency similar to those of previous studies were observed; however, when the tone was moved closer to the target segment assimilative effects were observed. In another experiment, contrastive effects of a following tone were observed in both CVC words and CV nonwords, although the size of the effects depended on syllable structure. Results are discussed with respect to contrastive mechanisms not speech-specific but operating at a relatively high level, taking into account spectrotemporal patterns occurring over extended periods before and after target events. © *2005 Acoustical Society of America.* [DOI: 10.1121/1.1984839]

PACS number(s): 43.71.An, 43.71.Es, 43.71.Pc [ALF] Pages: 1701–1710

I. INTRODUCTION

Human perception of speech sounds is heavily influenced by the acoustic properties of the contexts in which the sounds are heard. This context-dependence is of importance with respect to speech communication; due to coarticulatory influences from neighboring sounds, productions of a single phoneme may vary considerably across different phonological environments and perception must often take this variability into account in order to be effective. An oft-cited example of the reciprocal effects of context dependence in perception and production involves place of articulation in voiced stop consonants. The English phone [d] is distinguishable from the sound [g] primarily by a higher third formant (F3) just after the stop release. Due to coarticulation, however, this resonant frequency is a function not only of the intended consonant category but also of the speech sounds that neighbor it in a word or utterance. A [d] preceded by the liquid [r], for instance, tends to have a much lower, more "[g]-like" F3 than one preceded by [l]; conversely, a [g] preceded by [l] is more [d]-like (with a higher F3 frequency) than one following an [r]. Fortunately, however, these differences do not result in perceptual confusion, because listeners perceive the consonants relative to their environments. Mann (1980) observed that, given the same [d]-[g] series, listeners accept sounds with lower F3 values as instances of [d] when the sounds follow [r] as compared to [l], effectively compensating for the context-dependent differences in production of the consonants. Interestingly, listeners who are unlikely to rely on knowledge of English sound sequences also display this perceptual pattern. Japanese speakers, who do not perceptually distinguish between English [l] and [r]—and for whom liquid-stop sequences of any kind are phonotactically impossible—also identify the consonants in a context-dependent manner parallel to that shown by English speakers (Mann, 1986). Even 4-month-old infants discriminate [d]-like consonants from [g]-like ones at different F3 values depending on preceding liquids (Fowler *et al.*, 1990). Thus, listeners with limited or no phonetic knowledge nevertheless exhibit phonetic context effects. Effects of this nature are not limited to consonant identification, but seem to be ubiquitous in human speech perception. Across various places and manners of articulation [see Repp (1982) for a review], the pattern is the same: perception of speech sounds shows context influences which mirror—and thus tend to compensate for—the coarticulatory effects of neighboring sounds, allowing for accurate identification.

However, although context-dependent perception is of obvious linguistic benefit as one listens to continuous, naturally produced speech, it does not only occur under these circumstances. Listeners' perception of the same speech sounds is also influenced by nonspeech precursors such as pure tones. Lotto and Kluender (1998) observed that much like preceding [l] and [r] segments, frequency-modulated sine-wave glides approximating the [l] and [r] F3 trajectories and even steady-state F3 tones situated at the [l] and [r] F3 offset frequencies affect categorization of following [g]–[d] consonants. Importantly, the observed sine-wave-induced influence on speech categorization is contrastive in direction: lower-frequency ([r]-like) precursors cause more consonants to be identified as the relatively higher-frequency consonant alternative, [d], compared to higher-frequency ([l]-like) pre-

[a]Electronic mail: twade@andrew.cmu.edu

cursors. The directionality of the influence of nonspeech contexts is thus consistent with the effects of the speech contexts they model; coarticulation in speech is generally assimilative in nature and perceptual compensation is therefore generally in the opposite, contrastive direction. The observation that nonspeech stimuli that model very limited spectral characteristics of speech contexts produce similar effects of context, in conjunction with the finding that nonhuman animals also exhibit context-dependent identification of speech (Lotto *et al.*, 1997) has been taken to suggest that human context-dependent perception might rely on a general contrastive auditory mechanism rather than processes specialized for speech.

Considering Lotto and Kluender's (1998) finding that speech categorization was influenced by the spectral properties of an immediately preceding nonspeech sound, such a contrastive mechanism might be attributed to some relatively peripheral (cochlear or perhaps auditory nerve) process or processes. Early accounts of contrastive context-dependence proposed low-level processes such as peripheral neural adaptation as candidate mechanisms (Holt, 1999; Holt *et al.*, 2000) and forward masking has also been suggested as a possible cause for the nonspeech effects (Fowler *et al.*, 2000). However, it is unlikely that low-level sensory processes such as these are responsible for many of the context-dependent properties of speech perception. Phonetic context effects are observed even when context speech precedes a target speech sound by hundreds of milliseconds or is presented to the ear contralateral to that of speech target presentation (Holt and Lotto, 2002; Mann, 1980). Moreover, speech context may also influence phonetic categorization when it *follows* rather than precedes a speech target (Mann and Repp, 1980; Mann and Soli, 1991; Miller and Liberman, 1979). Mann and Repp (1980), for example, observed that although the frequency region of frication noise is a primary cue in distinguishing [s] from [š], fricative spectra also vary substantially depending on a following vowel. Fricatives followed by [u] have more energy at lower frequencies, and are therefore more [š]-like, than fricatives preceding [a]. However, as in the case of liquid-stop combinations, listeners tend to compensate for this effect perceptually, by accepting frication in lower frequency regions as signifying [s] in the presence of a following [u] compared to [a]. Since long-term, intra-aural, and backward-operating effects are unlikely to derive from masking or peripheral adaptation, it seems that context dependence in speech perception must be a product of higher-level processes involving more central mechanisms operating over longer time windows.

A majority of existing accounts assume that these high-level processes involve either some type of knowledge specific to speech (Gaskell, 2003; Gow, 2003; Nearey, 1997; Smits, 2001) or the perceptual recovery of articulatory events (Fowler, 1986; Fowler *et al.*, 2000; Liberman and Mattingly, 1985). These assumptions are called into question, however, by results suggesting that the influence of nonspeech precursors on speech categorization involves higher-level (nonperipheral) processing as well. Lotto *et al.* (2003) found that sine wave tones like those used by Lotto and Kluender (1998) influenced perception of following consonants even when nonspeech context and speech target were separated by up to 175 ms of silence or presented to opposite ears; these findings effectively rule out purely peripheral sensory mechanisms of masking and adaptation. Holt (2005) further observed that precursor "melodies" composed of multiple sine-wave tones similarly affected [g]-[d] categorization, even though the spectral characteristics of these nonspeech contexts were defined distributionally by sequential acoustic events that unfolded over seconds. When listeners heard a [g]-[d] series preceded by "high" and "low" sequences comprised of tones overlapping in absolute frequency but with means based on the [r] and [l] F3 values used by Lotto and Kluender (1998), "low" precursors robustly caused listeners to identify more members of a consonant series as [d], parallel to previous speech and nonspeech findings. Moreover, this nonspeech context effect persisted when as much as 1300 ms of silence or up to 13 intervening neutral (midfrequency) acoustic events separated the nonspeech precursors and the speech targets. Effects such as these are most consistent with higher-level (i.e., central, perhaps cortical) auditory processing and cannot be accounted for by solely peripheral mechanisms. Nor, however, can they be attributed to speech- or gesture-specific knowledge or abilities, since the context sounds involved were unambiguously synthetic. Thus, an emerging alternative explanation posits a *higher-level, nonlinguistic* mechanism whereby auditory events (either speech or nonspeech) occurring over time are incorporated into a general, contrast-providing context, relative to which targeted speech and other sounds are perceived (e.g., Diehl *et al.*, 2004; Holt, 2005; Wade and Holt, in press).

An important remaining asymmetry between speech and nonspeech context effects on speech categorization involves the directionality of the effects observed thus far. Whereas both preceding and following speech segments are known to affect phonetic categorization, documented effects of nonspeech contexts have been limited to the influence of precursors on following speech targets. The lack of data on backward-operating nonspeech contrastive effects leaves open the possibility that general auditory contrast may only operate in the forward direction. Conversely, accounts of perceptual compensation for coarticulation that invoke recovery of articulatory gestures (Fowler, 1986; Fowler *et al.*, 2000) or of temporally distributed acoustic features (Gow, 2004; Gow, 2003) readily account for patterns observed in the perception of both progressively and regressively coarticulated speech sounds. As a result, proponents of these accounts have described the general contrastive approach as unparsimonious in providing a mechanism for only a subset of the context effects observed in speech perception. Certainly, it is not necessarily the case that context dependence in the forward and backward directions arises from the same mechanism, even in speech perception. However, if the same central, high-level contrastive process reflected in the robust dichotic, temporally distributed influence of nonspeech precursors on speech observed in recent reports indeed contributes to compensation for carryover coarticulation in speech perception, it seems a reasonable hypothesis that the mechanism may also operate in the reverse direction, with later-occurring nonspeech contexts influencing perception of ear-

lier speech as well. The present study was designed to address this issue, testing whether spectral information from following nonspeech tones may effect the phonetic categorization of a preceding speech series.

In designing such a test, a major concern was that observation of backward effects of nonspeech context on speech perception may be constrained by methodological issues to a greater extent than observation of forward effects, perhaps accounting for the lack of data on the issue thus far. Specifically, we predicted that two methodological obstacles might relate to the way sequences of sounds are segregated into units and streams as a function of their temporal proximity (e.g., Bregman, 1990). First, insufficient temporal proximity between target speech and following nonspeech context might discourage listeners from considering the context in arriving at a speech categorization response. Since the contrastive influence of context-providing stimuli seems to be contingent on their perceptual continuity with target events (e.g., McCollough, 1965; Walker and Irion, 1979), it may be essential that listeners group the nonspeech events into a single stream with targeted speech sounds. This is especially critical when the context follows, rather than precedes, the target speech sound since the context can exert no effect if listeners make phonetic decisions before it is presented. It is known, for example, that following speech information has an attenuated contextual influence on the perception of earlier target speech sounds if listeners are not encouraged to take this information into account. Miller and Dexter (1988) found that although the total length of a syllable affects the classification of its onset consonant as voiced or voiceless, this influence is diminished if listeners respond very quickly after hearing the onset, suggesting that phonetic categorization decisions may be made without considering potentially informative following information. Under less speeded response conditions, Newman and Sawusch (1996) examined the extent to which temporally separated later-occurring speech context influenced the perception of word-initial consonants and found effects only when the context was within a short temporal window (a phoneme or two) after the target. Thus, we predicted that it would be necessary to place the nonspeech context events as closely as possible following target speech sounds to observe their effects.

A potentially competing issue, however, involves perceptual grouping at a lower hierarchical level. If the nonspeech context is presented too close to the target segment, it may be assimilated not only with the appropriate perceptual stream but with the target segment itself. That is, the spectral information provided by the nonspeech tone may be perceived as information *for* the speech target. Bregman (1990) has described processes of this type as fusion within a *unit* (as compared to a *stream*), and it may be similar or related to what occurs in duplex perception (e.g., Ciocca and Bregman, 1989; Liberman *et al.*, 1981) when a nonspeech chirp presented to the ear opposite an acoustically incomplete speech segment contributes to the perception of that segment. Should perceptual grouping along these lines occur, observable context effects of the nonspeech tones on speech categorization would be assimilative rather than contrastive in direction and therefore could override or obscure any effects of perceptual contrast that might occur.

In light of these methodological concerns, the present study attempts to strike a balance between providing for acoustic continuity such that nonspeech context is grouped into a context-providing stream with an earlier-occurring speech target while ensuring that the two are not perceived as information for a single event. Two experiments employed stimuli in which a brief nonspeech tone was inserted just after a target speech segment within a larger word context. The general reasoning behind this design was that a tone that was both preceded and followed by speech was more likely to be incorporated into a context-providing stream than a tone or sequence simply appended following a speech target. In addition, it was hypothesized that a word identification task requiring categorization of both initial and final consonants might further encourage listeners to take the nonspeech context into account; if acoustic information both preceding and following the nonspeech context were required for word recognition, listeners would be unable to respond (and therefore perhaps less likely to make a phonetic judgment on the target) before the nonspeech context sound occurred.

Experiment 1 tested for contrastive effects of following tones on target onset consonants in CVC words and also examined whether the opposite, assimilative, effects might occur if acoustic continuity (signaled by temporal proximity) between speech target and nonspeech context was sufficient. Experiment 2 tested the limits of this experimental design, examining whether any observed patterns would hold up in the absence of the lexical task or the CVC stimulus structure.

II. EXPERIMENT 1

Experiment 1 was designed as a first step in observing whether and how later-occurring nonspeech acoustic events may influence speech perception. Nonspeech context tones were embedded following the initial consonant in a CVC word stimulus. In keeping with previous studies (Holt, 2005; Lotto *et al.*, 2003; Lotto and Kluender, 1998; Mann, 1980), the initial (target) consonant was a stimulus drawn from a series of speech stimuli varying perceptually from [ga]-[da] and nonspeech contexts were pure tones of one of two frequencies shown in previous research to produce contrastive context effects on categorization of these [ga]-[da] series stimuli (Lotto and Kluender, 1998; Holt, 2005). In an effort to understand the situations in which perceptual grouping might lead to competing contrastive and assimilative effects on preceding speech, two conditions were tested, differing only in the temporal proximity of the embedded tone to the target consonant. It was hypothesized that nonspeech information occurring very close to the target consonant would be more likely to contribute to the perception of the consonant's spectral properties (e.g., Bregman, 1990; Wertheimer, 1923), whereas events occurring somewhat later would more likely contribute to a contrast-providing context. In Experiment 1a, nonspeech tones immediately followed the initial formant transitions, whereas in Experiment 1b they did not occur until well into the following vowel. To the extent that contrast effects occurred in either case, it was predicted that

listeners would label consonants [d] (the alternative with the higher F3 frequency) more often before a lower frequency tone than before a higher frequency tone. Assimilation, on the other hand, would result in more [d] responses in the high tone than in the low tone condition.

A. Method

1. Participants

Twenty-three college-age native English speakers from the Carnegie Mellon University community with no known speaking or hearing disorders participated in the study. Eleven participants were arbitrarily assigned to Experiment 1a and the remaining twelve participated in Experiment 1b. Participants were paid at least $7 per hour for their time.

2. Stimuli

Stimuli were English CVC words with pure tones replacing part of the vowel nucleus. Initial CV portions were drawn from a series varying perceptually from [ga] to [da] and the final consonant was an unambiguous [t] or [k], comprising the four-word set [dot, got, dock, gawk]. (For the experimenter, a speaker of North Midland American English, and most subjects interviewed, the vowel in *gawk* was identical to the [a] in the other three words. This vowel did not seem to present any difficulty identifying any of the words. Moreover, the acoustic characteristics of the vowel were identical across tokens, encouraging responses based only on the initial and final consonants.)

The consonant-vowel [ga]-[da] series was identical to that used by Holt (2005). The stimuli were derived from natural [da] and [ga] recordings from a monolingual male native English speaker (Computer Speech Laboratory; Kay Elemetrics, Lincoln Park, NJ; 20 kHz sample rate, 16 bit resolution), in the following manner. From a number of natural productions, one [ga] and one [da] token were selected that were nearly identical in spectral and temporal properties except for the onset frequencies of F2 and F3. LPC analysis was performed on each of the tokens, and a nine-step continuum of filters was created (Analysis-Synthesis Laboratory, Kay Elemetrics) such that the onset frequencies of F2 and F3 varied approximately linearly between [d] and [g] end points. These filters were excited by the LPC residual of the original [ga] production to create an acoustic series spanning the natural [da] and [ga] end points in approximately equal steps. The series was judged by the experimenters to comprise a gradual shift between natural-sounding [da] and [ga] tokens, and this impression was confirmed by regular shifts in phonetic categorization across the continuum by participants in the Holt (2005) study.

For the present experiment, these CV segments were trimmed to 275 ms in length from the initial burst and their amplitudes were attenuated linearly to zero over the final 50 ms. They were followed by a 25 ms silent interval, representing a stop closure, and 170 ms of burst from a [t] or [k] taken from recordings of natural native-English male productions of the consonants in monosyllabic word-final position following a low back vowel. The resulting stimuli were judged to resemble careful natural productions of the words *dot, dock, got*, and *gawk*.

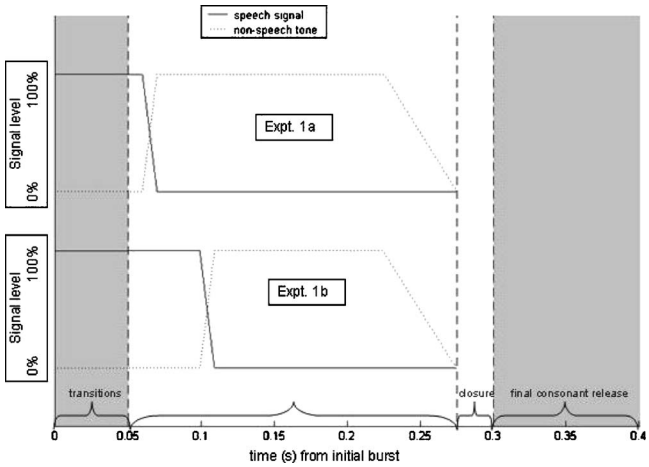

FIG. 1. Schematic illustration of speech and nonspeech stimuli in Experiment 1 conditions.

Figure 1 shows the composition of Experiment 1 stimuli. In two conditions, nonspeech sine-wave tones were inserted to replace the final portion of the vowel. In Experiment 1a, tones were introduced immediately after the initial formant transitions (transitions lasted approximately 50 ms following the initial burst), replacing the final 215 ms of the vowel. In Experiment 1b tones did not occur until further into the vowel, replacing only the final 175 ms. Tone insertion was achieved as follows: beginning either 60 ms (Experiment 1a) or 100 ms (Experiment 1b) after an initial burst, a tone was gradually introduced over the duration of the vowel, so that the vowel:tone amplitude ratio decreased linearly to zero over 10 ms. The vowel was completely replaced by the tone over the remaining 205 ms (Experiment 1a) or 165 ms (Experiment 1b), with a linear amplitude off-ramp over the final 50 ms. The resulting stimuli gave the impression of spoken word recordings to which synthetic beeps had been added. Speech onset, nucleus, and offset portions were identifiable in all stimuli, although the vowel was necessarily more ambiguous in Experiment 1a stimuli.

Tone frequencies were selected to provide spectral contrast for the initial consonant F3 frequencies. Following Lotto and Kluender (1998), the sine-wave tones had a periodicity of either 1800 or 2800 Hz.

3. Procedure

Participants heard each [ga]-[da] stimulus series member with each of the embedded tone conditions [2800 Hz tone, 1800 Hz tone] and each final consonant 18 times; 18 additional repetitions of each CVC combination without embedded tones were included as filler items, for a total of 972 stimuli. Presentation was controlled by Tucker Davis Technologies (TDT) System II hardware; stimuli were converted from digital to analog, low-pass filtered at 4.8 Hz, amplified, and presented diotically over linear headphones (Beyer DT-150) at approximately 70 dB SPL to participants in sound-attenuated booths. Stimuli were presented in random order in

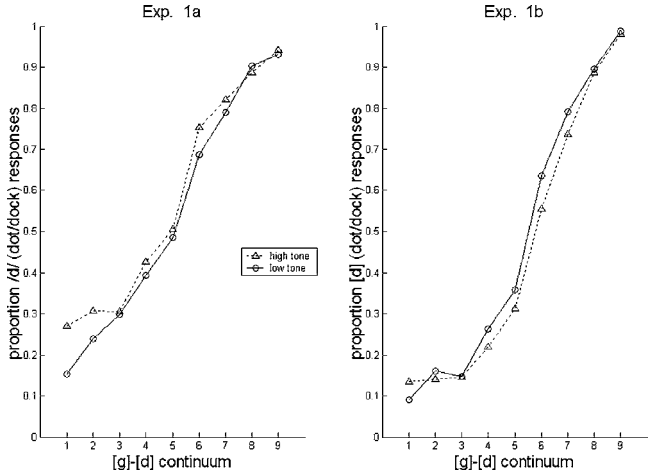

FIG. 2. Initial consonant classification in word response patterns for Experiment 1 conditions, averaged over offset consonants and presented as the proportion of [d] responses across CV stimuli.

two identical sessions separated by a short break; participants were instructed to listen to each word and press one of four buttons labeled *got, dot, gawk, dock* in response.

B. Results

Figure 2 shows listener responses to the [ga]-[da] continua for the two tone conditions, averaged across final consonants. No-tone filler items were not included in the analysis, since interpretable hypotheses focused on context effect differences depending on tone frequencies. Overall, Experiment 1a participants were somewhat more variable in responding, with [d] and especially [g] end points identified at less than 100%. This was not unexpected, since Experiment 1a stimuli lacked speech information immediately following the onset formant transitions that was present in Experiment 1b.

1. Context influence on target consonants

In Experiment 1a, a reliable effect of tone frequency was observed; participants responded [d] (*dot* or *dock*) more often when the higher frequency tone immediately followed the initial consonant than when the lower frequency tone followed it, $t(10)=3.14$, $p=0.011$. This indicates that the immediately adjacent tone contexts of Experiment 1a influenced perception of the preceding consonants assimilatively. Spectral energy in a *higher* frequency range resulted in more [d] (i.e., *high*-F3) responses. Experiment 1b participants also showed a reliable overall effect of the following context tone frequency on categorization, $t(11)=3.81$, $p=0.003$. However, the effect in this case, as in previous studies involving preceding speech (Mann, 1980), preceding nonspeech (e.g., Lotto and Kluender, 1998) and following speech (Mann and Repp, 1980) contexts, was *contrastive* in direction. The *lower*-frequency tone caused listeners to categorize the consonants more often as [da], the alternative with the *higher* F3 frequency.

Thus, later-occurring nonspeech may influence speech categorization, both assimilatively and contrastively. Moreover, it seems that a shift between these two influences (or a shift in the effectiveness with which one or both of them operates) may be induced by a change of as little as 40 ms in the asynchrony of the nonspeech context with respect to the target speech segments. To confirm this shift in the present between-subjects design, the effects of the nonspeech contexts were compared across participant groups. A 2(tone frequency)$\times$2(tone-onset, Experiments 1a, 1b) mixed-model ANOVA with tone-onset condition as a between-subjects factor revealed a significant interaction, $F(1,21)=21.96$, $p<0.001$, indicating that tone frequency indeed had a differential effect across the two groups as a function of the temporal proximity of the tone to the speech target.

In speculating on the processes driving these context effects, their magnitude (the mean difference in [d] responses across tone conditions) was compared with participants' median reaction times. For Experiment 1a, for which assimilative context effects of temporally adjacent tones were observed, bivariate (Pearson) correlation revealed no relationship between context effect size and reaction time, $p>0.5$. For Experiment 1b where tone onset was somewhat later in the vowel, however, a reliable positive correlation was observed, $r=0.606$, $p=0.037$. Participants who took longer to respond generally showed larger contrastive effects of the following tone on speech categorization. This pattern may simply have been due to a subset of participants whose fast, careless responses led to decreased effect size; however, investigation revealed no similar correlation between response time and any measure of accuracy or consistency. Possible implications of this observation with respect to the workings of a high-level contrastive perceptual mechanism are addressed in Sec. IV.

2. Final consonant effects

Since this experiment involved word decisions that depended on final consonants as well as initial, target consonants, and since the word-final release burst stimuli used had not been previously tested experimentally, we also examined listener responses with respect to these consonants. Listeners responded consistently with the intended consonant (e.g., *dot* or *got* for a final [t] burst) on 94% of trials, indicating that the consonants were indeed unambiguous. This observation was further confirmed by the fact that no effect of tone frequency on final consonant identification was observed in either test condition, $(t)<1$.

A 2(final consonant)$\times$2(tone frequency)$\times$2(condition) mixed model ANOVA with proportion [d] responses as the dependent variable revealed a main effect of final consonant [$F(1,21)=25.5$; $p<0.001$] but no interactions involving the consonant, $F<1$, indicating that listeners robustly categorized the initial consonant as [d] more often when the word ended in [k] than when it ended in [t]. The present results cannot conclusively determine the source of this effect, but it is not surprising. One possibility is bias deriving from word frequency (Connine *et al.*, 1993). While frequency measures are not available for the word *gawk*, it is almost certainly many times less common than any of the other words [Kucera-Francis (1967) written frequencies for *dot, dock*, and *got* are 13, 8, and 482, respectively] which may have

biased listeners disproportionately against this word. Alternatively, the final consonant may have been exerting additional spectrally contrastive perceptual effects on the onset, since the final [k] spectra were probably more similar to those of initial [g] than [d] consonants. In any case, this effect did not seem to interfere in any way with the main focus of the experiment, the effects of embedded tones on onset consonants.

III. EXPERIMENT 2

Experiment 1 demonstrated the existence of backward-operating contrastive influence of nonspeech on speech and addressed one of the two methodological issues that were raised concerning the observation of this influence, namely whether contrastive effects may be obscured by an assimilative effect as a function of temporal proximity to target speech. The results demonstrated that context effects produced by nonspeech context tones that follow a speech target are modulated by the temporal proximity of the context and target sounds. Listeners may assimilate the nonspeech sound as information for the speech target when it follows the speech target immediately, but perceive it in contrast to the target when it occurs somewhat later.

The other potential methodological issue involved ensuring that the acoustic continuity between target and following context was sufficient to drive the contrastive effect seen in Experiment 1b. Experiment 1 showed that the backward nonspeech effects may be observed in certain favorable circumstances where (1) additional speech information follows the nonspeech event and (2) this information is required for a proper response. To determine whether these stimulus and task construction characteristics are essential in observing context effects of later-occurring nonspeech sounds on speech categorization and in investigating the limits of the context effects observed in Experiment 1, an additional study was designed in which the final consonant present in Experiment 1 was either eliminated or was not critical to the required categorization response.

A. Methods

1. Participants

Participants were 16 college-age native English speakers with no known speaking or hearing disorders. Participants were paid at least $7 per hour for their time.

2. Stimuli

There were two conditions tested in Experiment 2. Stimuli for Condition 1 were English CVC words and embedded tones identical to those for which contrastive effects were observed in Experiment 1b. Condition 2 stimuli were CV syllables created by eliminating the final-consonant burst of the Experiment 1b stimuli, leaving only the consonant-vowel and embedded tones.

3. Procedure

Participants heard 486 Condition 1 stimuli [(2 tone conditions+no-tone fillers)$\times$2 final consonants$\times$9 continuum members$\times$9 repetitions] and 486 Condition 2 stimuli in consecutive sessions separated by a short break. Half of the participants heard Condition 1 stimuli in the first session and Condition 2 stimuli in the second, and the other half heard Condition 2 stimuli first. Stimulus presentation was identical to that of Experiment 1. Rather than making a word decision, however, in both conditions participants were instructed simply to quickly press one of two buttons depending on the first consonant (*d* or *g*) of a stimulus.

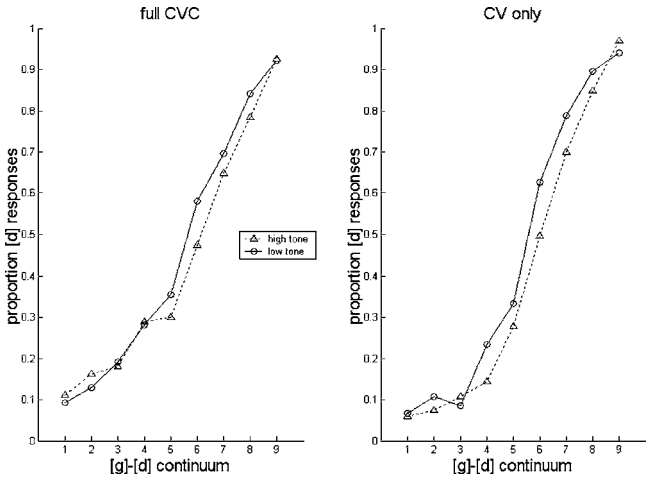

FIG. 3. Initial consonant classification in Experiment 2 conditions.

B. Results

One participant displayed an irregular, apparently random response pattern; this participant's data were discarded. Figure 2 shows response curves to the initial [g]-[d] continuum as a function of the embedded tone for the remaining 15 participants. As in Experiment 1b, tone frequency reliably influenced consonant identification in a contrastive manner; participants more often reported hearing an initial [d] in the presence of the lower-frequency (1800 Hz) tone in both CVC, $t(15)=2.62$, $p=.02$, and CV, $t(14)=3.95$, $p=0.001$, syllables. For neither condition did the size of the effect $[p([\mathrm{d}],\text{low tone})-p([\mathrm{d}],\text{high tone})]$ differ reliably from that of Experiment 1b, indicating that neither the CVC word context nor the lexical task requiring the final syllable in responses was essential in driving the effect.

Figure 3 shows a comparison of the effect sizes across Experiment 2 conditions and session orders. A 2(condition, syllable type)$\times$2(session order) mixed model ANOVA with session order as a between-subjects factor revealed that participants demonstrated slightly larger contrastive effects overall when no final consonant was present, $F(1,13)=8.83$, $p=0.011$. Additionally, across conditions a robustly larger contrastive effect was seen in the participants' second sessions, without respect to whether the stimuli were CVCs or CVs, resulting in a Condition$\times$Session Order interaction ($F(1,13)=21.8$, $p<0.001$). This does not seem to have been an artifact of the task or an effect of practice. In Experiment 1b, which was identical to Experiment 2 except for the response choices and the shift in syllable structure between sessions, there was a nonsignificant shift in the opposite direction, with participants showing

greater effects in the first session. A 2(Experiment 1b versus 2) × 2(Session) mixed model ANOVA with effect size as the dependent variable revealed a significant interaction, $F(1,25)=9.35$, $p=0.005$, confirming that a session-to-session pattern different from Experiment 1b emerged in Experiment 2. Implications of this pattern are discussed in the following section.

IV. GENERAL DISCUSSION

The present study demonstrates that later-occurring nonspeech acoustic events can influence categorization of preceding speech. In two experiments, the periodicity of brief pure tones inserted after members of a target [g]-[d] continuum reliably affected listeners' speech identification. Like previously documented effects of preceding speech (Mann, 1980) and nonspeech (Holt, 2005; Lotto and Kluender, 1998) on speech perception, of preceding speech on nonspeech perception (Stephens and Holt, 2003), and following speech on speech perception (Mann and Repp, 1980), this influence was primarily contrastive in nature. Higher-frequency following tones caused ambiguous speech targets to be identified as the lower-F3 consonant alternative ([g]), and *vice versa*.

This finding adds to the array of findings in which nonspeech acoustic context has been shown to parallel speech context in its perceptual influence on speech. As such, it is consistent with the view that a general contrastive perceptual mechanism may play a role in driving the context-dependent nature of speech perception and calls into question the extent to which specialized knowledge or processes specific to articulation are needed to explain context-dependent speech perception. In particular, demonstration that nonspeech context effects are not constrained in the temporal direction of their operation increases the parsimony with which a contrast-based explanation can account for the set of observed speech context effects. Like feature parsing (Gow, 2004; Gow, 2003) and gesture recovery (Fowler, 1986; Fowler *et al.*, 2000) accounts, contrast effects could potentially explain observed perceptual compensation for both progressively and regressively coarticulated speech. This is not, however, to say that the contrastive effect demonstrated here is the only factor contributing to the context-dependent nature of speech perception. As is often noted in conjunction with proposals of contrastive effects, it seems likely that speech- or language-specific learning also plays a role (Diehl *et al.*, 2004; Holt and Kluender, 2000; Lotto, 2000). Since contrast effects do not arise as compensation for speech production, it is unreasonable to predict that they would result in ideal perceptual compensation for all possible patterns of coarticulation, which probably vary somewhat from language to language. Although general contrastive effects probably simplify the perceptual problem by perceptually eliminating some context-conditioned spectral variability, residual variability might be learned as part of language acquisition. Indeed, there is evidence that speakers of different languages process context-dependent variability differently. Beddor, Harnsberger, and Lindemann (2002), for example, observe cross-language differences in perceptual compensation for vowel coarticulation, citing differences between English and Shona speakers' category boundary shifts depending on the position of context-providing vowels.

The results of Experiment 1 also demonstrate the sensitivity of nonspeech-induced context effects to the spectrotemporal characteristics of the stimuli. A robust shift from assimilative to contrastive effects resulted from a 40 ms change in the temporal proximity of the nonspeech context to the target. This observation underscores the variety of ways nonspeech (or speech) context sounds might influence speech perception and also suggests some caution in the interpretation of null effects of nonspeech context (e.g. Fowler *et al.*, 2000), since such observations might stem from the operation of multiple, competing perceptual effects like those observed here.

A. Temporal range of context effects

A principle concern in designing Experiment 1b and Experiment 2 stimuli was to provide sufficient temporal proximity between target and context stimuli for contrastive processes to operate. We did observe contrastive influences when the context was placed some 40 ms after the target in the following vowel segment, but the present experiments do not address whether these influences might persist if the context were further separated in time. Newman and Sawusch (1996) observed that perception of a segmental contrast involving duration may only be affected by speaking rate context information that follows the target segment by no more than a phoneme or two. A parallel study involving the limits of following nonspeech context will be informative in further comparing the perceptual influences of speech and nonspeech contexts.

B. Contrast and higher levels of linguistic processing

These results and those of previous studies involving nonspeech context effects on speech categorization (e.g., Holt, 2005; Lotto *et al.*, 2003; Wade and Holt, in press) suggest that a *higher-level* (central), *nonlinguistic* contrastive mechanism may play a role in the interpretation of speech sounds. However, we wish to emphasize that such a mechanism need not be independent of, or incompatible with, the various language-specific processes involved in human speech perception. In particular, the present results resonate with spoken word recognition models that posit interaction between multiple representational levels over the course of perception. In the TRACE model (McClelland and Elman, 1986), for example, it is assumed that speech processing takes the form of activation at three hierarchically arranged levels: acoustic features, abstract phonemes, and words. Although the initial input to the model involves the feature units, there are bidirectional excitatory connections between units at neighboring levels; critically, the units at all levels corresponding to all of the segments in a word remain active—and continue to interact—for the entire time a word is processed. While we will not argue that speech processing involves precisely these three levels of representation, it follows straightforwardly from the interactive dynamics of the model that if perception of the acoustic events comprising

speech sequences is assumed to be modulated by the acoustic context—speech or nonspeech—preceding and following the events, the unit activations at lower representational levels might adjust over time as context becomes available. As input representations evolve over time, the activation of the higher-level units they excite should also change, although the effects of the change would be delayed by factors including the strength of the top-down connections of higher-level units already activated by the precontext feature representations. Thus, assuming a general auditory representation as a lower-level input representation to an interactive model like TRACE provides a means by which a general contrastive mechanism might interact with higher-level, language-specific processes over the course of perception.[1]

Experiment 1b revealed a pattern in listeners' reliance on nonspeech context that might shed some light on the dynamics of such interaction. As response latencies increased, so did the magnitude of the contrastive effect of following tone frequency on their speech categorization. While care should be taken not to over-interpret this correlational observation, it is consistent with the notion that top-down influences delay the effects of acoustic context at the input level. Since this process might continue well beyond the physical duration of the word, delaying word recognition (presumably reflected in the longer response times) should allow the context to exert a greater contrastive effect.

The results of Experiment 2 demonstrate the robustness of the backward contrastive effect and provide some additional evidence regarding the types of processing it might involve. In two conditions where participants were asked to label only the initial consonants of stimuli as [d] or [g], later-occurring tones identical to those used in Experiment 1b evoked similar contrastive influences on categorization. This indicates that the CVC word identification task, which required participants to make a (word) response based on parts of a speech signal both preceding and following the nonspeech tone, was in fact not critical in producing the backward effect. Additionally, participants showed a slightly greater contrastive effect when the stimuli were CV nonwords than when they were CVC words. The cause of this difference is not clear; it may have been that the additional later-occurring context provided by the final consonant actually obscured the contrastive effect of the tones. Alternatively, the lexical status of the stimuli may have played a role. It has been observed that listeners are faster at phoneme monitoring in word than in nonword contexts (Cutler *et al.*, 1987; Eimas *et al.*, 1990), though not always with stimuli as short as those used here (Foss and Gernsbacher, 1983). If, as suggested by these authors, word stimuli resulted in a different mode of processing in for the words in Condition 1 than the nonwords in Condition 2, the resulting faster processing of onsets in word stimuli might have resulted in less contrastive influence of later-occurring context. Reaction time comparisons in this experiment were confounded by the additional differences in syllable length and structure (CV vs CVC) across conditions and were thus uninformative in this respect.

Comparison of effects across Experiment 2 conditions also revealed an interaction involving session order. Regardless of syllable structure, participants exhibited reliably greater contrastive effects during their second test sessions, an effect not observed in Experiment 1. [See Fig. 4.] While further study will be needed to determine the precise causes of this difference, it seems likely that attention played a role. Second test session stimuli always involved novel word-final information compared to those of the first session (either the addition or the sudden absence of a coda consonant); it seems likely that this novelty resulted in increased attention to portions of the stimulus following the initial onset, perhaps including the nonspeech tones. This increased attention, then, might have enhanced the tones' contrastive influence on the onset consonants, resulting in the session-to-session effect-size difference seen in Experiment 2.

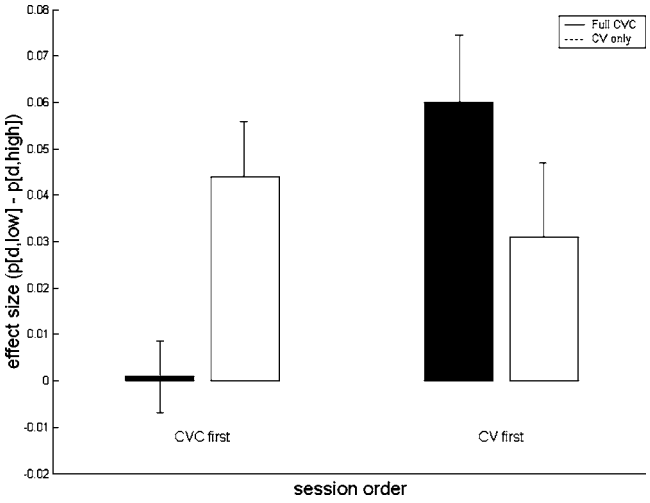

FIG. 4. Experiment 2 contrastive effect sizes across syllable types and session orders (error bars indicate standard errors).

C. Masking as a possible account

Effects of preceding nonspeech acoustic context on speech categorization have previously been attributed to auditory masking (Fowler *et al.*, 2000), although this interpretation is challenged by the extended time courses and dichotic presentation paradigms (Holt, 2005; Lotto *et al.*, 2003) for which the nonspeech context effects have been observed. Holt (2005), for example, presents evidence that nonspeech precursor contexts influence speech perception even when 13 constant acoustic stimuli spanning 1.3 s intervene between nonspeech context and speech targets. Masking is very unlikely to account for such context effects of nonspeech on speech. Nonetheless, the backward-operating effects documented here might similarly be attributed to a process such as backward masking (e.g., Tyler and Small, 1977) or informational masking (Neff *et al.*, 1993; Pollack, 1975). By selectively masking spectral energy in the neighborhood of 2800 Hz, for example, a higher-frequency following tone could cause an ambiguous consonant to be perceived as more [g]-like, resulting in contrastive effects as reported here. The present experiments do not rule out these possibilities; indeed, given the relatively poor current understanding of the precise mechanisms responsible for backward and informational masking, it is difficult to determine

whether context-dependence in natural speech might derive from similar or related processes. Backward masking, for example, has been observed to operate even when the masking stimulus is presented to the ear contralateral to the target (Weber and Green, 1979), indicating a more central mechanism. In fact, backward masking has been mentioned as a potential contributor to the context-dependent perception of various speech contrasts (Dent *et al.*, 1997; Jamieson, 1987; Sinnott *et al.*, 1998). Further research and a better understanding of these processes will be required to determine whether this is indeed the case.

D. Conclusions

This study demonstrates that later-occurring nonspeech events contrastively influence the categorization of preceding speech sounds in certain circumstances. Contrastive backward effects of pure tones on onset consonant identification were observed across syllable structures (CV and CVC) and tasks (word versus phoneme identification), although they depended on the temporal proximity of nonspeech tones to the preceding speech target and perhaps on attention to later portions of the stimuli. These findings are taken as support of a central, nonlinguistic contrastive perceptual mechanism, whereby auditory events occurring over time are incorporated into an acoustic context, relative to which preceding or following speech sounds are perceived.

ACKNOWLEDGMENTS

This work was supported by a James S. McDonnell Foundation award for Bridging Mind, Brain, and Behavior to LLH, NIH Grant No. 5 RO1 DC04674-02 to L.L.H., and by a fellowship from the NIH Postdoctoral Training Grant on "Individual Differences in Cognition." The authors thank Christi Adams and Ashley Episcopo for help in conducting the experiments.

[1]The original TRACE model, it should be noted, has a different, phonologically oriented account of context-dependent perception which does not account for nonspeech context effects (McClelland and Elman, 1986).

Beddor, P. S., Harnsberger, J. D., and Lindemann, S. (**2002**). "Language-specific patterns of vowel-to-vowel coarticulation: Acoustic structures and their perceptual correlates," J. Phonetics **30**, 591–627.

Bregman, A. S. (**1990**). *Auditory Scene Analysis* (MIT, Cambridge, MA).

Ciocca, V. and Bregman, A. S. (**1989**). "The effects of auditory streaming on duplex perception," Percept. Psychophys. **46**, 39–48.

Connine, C. M., Titone, D., and Wang, J. (**1993**). "Auditory word recognition: Extrinsic and intrinsic effects of word frequency," Cognition **19**, 81–94.

Cutler, A., Mehler, J., Norris, D., and Segui, J. (**1987**). "Phoneme identification and the lexicon," Cogn. Psychol. **19**, 141–177.

Dent, M. L., Brittan-Powell, E. F., Dooling, R. J., and Pierce, A. (**1997**). "Perception of synthetic /ba/-/wa/ speech continuum by budgerigars (*Melopsittacus undulatus*)," J. Acoust. Soc. Am. **102**, 1891–1897.

Diehl, R., Lotto, A., and Holt, L. L. (**2004**). "Speech perception," Annu. Rev. Psychol. **55**, 149–179.

Eimas, P., Hornstein, S. B., and Payton, P. (**1990**). "Attention and the role of dual codes in phoneme monitoring," J. Mem. Lang. **31**, 375–395.

Foss, D. J. and Gernsbacher, M. A. (**1983**). "Cracking the dual code: Toward a unitary model of phoneme identification," J. Verbal Learn. Verbal Behav. **22**, 609–632.

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

Fowler, C. A., Brown, J. M., and Mann, V. A. (**2000**). "Contrast effects do not underlie effects of preceding liquids on stop-consonant identification by humans," J. Exp. Psychol. Hum. Percept. Perform. **26**, 877–888.

Gaskell, M. (**2003**). "Modelling regressive and progressive effects of assimilation in speech perception," J. Phonetics **31**, 447–463.

Gow, D. (**2004**). "A cross-linguistic examination of assimilation context effects," J. Mem. Lang. **51**, 279–296.

Gow, D. W. (**2003**). "Feature parsing: Feature cue mapping in spoken word recognition," Percept. Psychophys. **65**, 575–590.

Holt, L. L. (**1999**). "Auditory constraints on speech perception: An examination of spectral contrast," Diss. Abstr. Int., C **61**, 556.

Holt, L. L. (**2005**). "Temporally non-adjacent non-linguistic sounds affect speech categorization," Psychol. Sci. **16**, 305–312.

Holt, L. L. and Kluender, K. R. (**2000**). "General auditory processes contribute to perceptual accommodation of coarticulation," Phonetica **57**, 170–180.

Holt, L. L. and Lotto, A. J. (**2002**). "Behavioral examinations of the level of auditory processing of speech context effects," Hear. Res. **167**, 156–169.

Holt, L. L., Lotto, A. J., and Kluender, K. R. (**2000**). "Neighboring spectral content influences vowel identification," J. Acoust. Soc. Am. **108**, 710–722.

Jamieson, D. (**1987**). "Studies of possible psychoacoustic factors underlying speech perception," in *The Psychophysics of Speech Perception*, edited by M. E. H. Schouten (Nijhoff, Dordrecht).

Kucera, H. and Francis, W. N. (**1967**). *Computational Analysis of Present-day American English* (Brown University Press, Providence).

Liberman, A., Isenberg, D., and Rakerd, B. (**1981**). "Duplex perception of cues for stop consonants: Evidence for a phonetic mode," Percept. Psychophys. **30**, 133–143.

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

Lotto, A., Sullivan, S., and Holt, L. L. (**2003**). "Central locus for nonspeech context effects on phonetic identification," J. Acoust. Soc. Am. **113**, 53–56.

Lotto, A. J. (**2000**). "Language acquisition as complex category formation," Phonetica **57**, 189–196.

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

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.

Mann, V. A. and Soli, S. (**1991**). "Perceptual order and the effect of vocalic context on fricative perception," Percept. Psychophys. **49**, 399–411.

McClelland, J. L. and Elman, J. L. (**1986**). "The TRACE model of speech perception," Cogn. Psychol. **18**, 1–86.

McCollough, C. (**1965**). "Color adaptation of edge-detectors in the human visual system," Science **149**, 1115–1116.

Miller, J. L. and Dexter, E. R. (**1988**). "Effects of speaking rate and lexical status on phonetic perception," J. Exp. Psychol. Hum. Percept. Perform. **14**, 369–378.

Miller, J. L. and Liberman, A. M. (**1979**). "Some effects of later-occurring information on the perception of stop consonant and semivowel," Percept. Psychophys. **25**, 457–465.

Nearey, T. M. (**1997**). "Speech perception as pattern recognition," J. Acoust. Soc. Am. **101**, 3241–3254.

Neff, D. L., Dethlefs, T. M., and Jesteadt, W. (**1993**). "Informational masking for multicomponent maskers with spectral gaps," J. Acoust. Soc. Am. **94**, 3112–3126.

Newman, R. S. and Sawusch, J. R. (**1996**). "Perceptual normalization for speaking rate: Effects of temporal distance," Percept. Psychophys. **58**, 540–560.

Pollack, I. (**1975**). "Auditory informational masking," J. Acoust. Soc. Am. **57**, S5.

Repp, B. H. (**1982**). "Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception," Psychol. Bull. **92**(1), 81-110.

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.

Smits, R. (**2001**). "Evidence for hierarchical categorization of coarticulated phonemes," J. Exp. Psychol. Hum. Percept. Perform. **27**, 1145–1162.

Stephens, J. D. W. and Holt, L. L. (**2003**). "Preceding phonetic context affects perception of non-speech sounds," J. Acoust. Soc. Am. **114**, 3036–3039.

Tyler, R. S. and Small, A. M. (**1977**). "Two-tone suppression in backward masking," J. Acoust. Soc. Am. **62**, 215–218.

Wade, T. and Holt, L. L. "Perceptual effects of preceding non-speech rate on temporal properties of speech categories," Percept. Psychophys. (in press).

Walker, J. T. and Irion, A. L. (**1979**). "Two new contingent aftereffects: Perceived auditory duration contingent on pitch and on temporal order," Percept. Psychophys. **26**, 241–244.

Weber, D. L. and Green, D. M. (**1979**). "Suppression effects in backward and forward masking," J. Acoust. Soc. Am. **65**, 1258–1267.

Wertheimer, M. (**1923**). "Untersuchungen zur Lehre der Gestalt" ("Investigations of Gestalt Theory"), Psychol. Forsch. **4**, 301–350.