Central locus for nonspeech context effects on phonetic identification (L)

Andrew J. Lotto[a] and Sarah C. Sullivan[b]
Department of Psychology, Washington State University, P.O. Box 644820, Pullman, Washington 99164

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

(Received 29 May 2002; revised 9 October 2002; accepted 14 October 2002)

Recently, Holt and Lotto [Hear. Res. **167**, 156–169 (2002)] reported that preceding speech sounds can influence phonetic identification of a target syllable even when the context sounds are presented to the opposite ear or when there is a long intervening silence. These results led them to conclude that phonetic context effects are mostly due to nonperipheral auditory interactions. In the present paper, similar presentation manipulations were made with nonspeech context sounds. The results agree qualitatively with the results for speech contexts. Taken together, these findings suggest that the same nonperipheral mechanisms may be responsible for effects of both speech and nonspeech context on phonetic identification.
[DOI: 10.1121/1.1527959]

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

I. INTRODUCTION

There exists a class of perceptual phenomena known as *phonetic context effects* in which the perceived phonemic identity of a speech sound is moderated by the identity of neighboring speech sounds. That is, identical acoustics can lead to different identifications depending on the identity of precursor speech sounds. For example, the reported identification of a syllable-initial stop can be changed from /g/ to /d/ by changing the preceding context syllable from /al/ to /ar/ (Mann, 1980).

Holt and Lotto (2002) attempted to ascertain the level of the auditory system at which the stimulus interactions underlying phonetic context effects occur. In one experiment, they presented context syllables (e.g., /al/ or /ar/) and target syllables (/da/–/ga/ series members) to opposite ears. The identity of the context syllable affected identifications of the target syllable even in this dichotic presentation condition. However, the size of the identification boundary shift was slightly smaller than for diotic presentation conditions. In a second experiment the duration of the silent gap between the context and target syllable was varied from 25 to 400 ms (this gap was typically 50 ms in previous experiments). A significant effect of context was evident even when context offset and target onset were separated by as much as 275 ms. Holt and Lotto (2002) argue that these results suggest that context effects are partially mediated by nonperipheral mechanisms. That is, it is unlikely that they are due to masking or interactions at the level of the auditory nerve or perhaps even cochlear nucleus. In agreement with these conclusions, Holt and Rhode (2000) failed to find evidence for appropriate speech-sound stimulus interactions in recordings from chinchilla VIIIth nerve.

Recently, there have been a number of demonstrations of shifts in phonetic identification caused by nonspeech context sounds such as sine-wave tones (Lotto and Kluender, 1998; Holt *et al.*, 2000). Lotto and Kluender (1998) presented listeners consonant–vowel (CV) syllables preceded by sine-wave tones that modeled the frequency trajectory of the third formant ($F3$) of /al/ or /ar/. Listeners identified the CVs more often as /ga/ following the sine-wave modeling /al/ and more often as /da/ following the sine-wave modeling /ar/. Because these nonspeech context sounds had no perceived phonetic content, the authors proposed that the spectral content of the context sounds moderates the shift in identity of the target speech sounds. In this case, high-frequency spectral energy ($F3$ offset of /al/ or high-frequency sine wave) leads to more /ga/ responses (/g/ has a low-frequency $F3$ onset) and low-frequency spectral energy ($F3$ offset of /ar/ or low-frequency sine wave) results in more /da/ responses (/d/ has a high-frequency $F3$ onset). This pattern of results has been referred to as *spectral contrast* (Holt *et al.*, 2000).

The question that is immediately raised is whether the processes responsible for nonspeech context effects are the same as those underlying speech context effects. Fowler *et al.* (2000) suggest that nonspeech context effects are primarily due to masking. On the other hand, they propose that speech context effects are due specifically to perception of speech gestures.

In addition to a masking account, it is possible that nonspeech context effects are complex demonstrations of *auditory enhancement* (Viemeister, 1980; Viemeister and Bacon, 1982; Summerfield *et al.*, 1984). Auditory enhancement refers, generally, to a class of effects in which energy in a frequency region is perceptually enhanced if it is preceded by a sound that lacks energy in that region.

Holt and Lotto (2002) argue that their results are incom-

[a]Electronic mail: alotto@wsu.edu
[b]Sarah Sullivan is currently in the Department of Psychology, University of Texas-Austin.

patible with an auditory enhancement or peripheral masking account of *speech* context effects. In particular, the time course of auditory enhancement appears to differ from the speech context effects. Holt and Lotto demonstrated that speech context effects are present out to at least 275 ms of intervening silent gap. Viemeister and Bacon (1982) found no appreciable auditory enhancement in a masking study beyond about 100 ms of intervening silence. In addition, auditory enhancement appears to be a strictly monaural phenomenon. Summerfield and Assmann (1989) failed to find effects of a precursor stimulus in auditory enhancement vowel experiments when the context was presented to the contralateral ear. In contrast, Holt and Lotto demonstrated robust effects of speech contexts presented to the opposite ear of the target syllables.

The purpose of the two experiments presented here is to determine whether auditory enhancement or peripheral masking can completely account for nonspeech context effects. The manipulations utilized by Holt and Lotto (2002) have been replicated here with nonspeech contextual sounds. Experiment 1 examines the effect of dichotic versus diotic presentation on nonspeech context effects. In experiment 2, subjects are presented context and target syllables with varying durations of intervening silence. The question is whether these manipulations will moderate nonspeech context effects in a qualitatively different manner than witnessed for speech context effects. If not, then it may be reasonable to suggest that similar mechanisms are culpable for both speech and nonspeech context effects.

II. EXPERIMENT 1 (DICHOTIC VERSUS DIOTIC PRESENTATION)

A. Methods

1. Subjects

Twenty-four undergraduate students at Washington State University participated in the experiment for course credit. All were native English speakers that reported no hearing deficits or disorders.

2. Stimuli

A ten-member series of synthetic speech varying acoustically in $F3$ onset frequency and varying perceptually from /ga/ to /da/ was created using the cascade branch of the Klatt (1980) synthesizer. For these stimuli, $F3$ onset frequency varied from 1800 to 2700 Hz in 100-Hz steps. From onset, $F3$ frequency changed linearly to a steady-state value of 2450 Hz across 80 ms. All other synthesis parameters were constant across series members. The first formant frequency ($F1$) increased linearly from 300 to 750 Hz and the second formant ($F2$) frequency declined from 1650 to 1200 Hz across 80 ms. The fourth formant ($F4$) had a steady-state value of 2850 Hz. Fundamental frequency ($f0$) was 110 Hz over the first 200 ms and decreased to 95 Hz over the last 50 ms. Total stimulus duration was 250 ms. This CV series is identical to that used by Holt and Lotto (2002) in experiments 1b and 2b.

The nonspeech context stimuli were based on the speech precursors used in Holt and Lotto (2002). An analog of /al/ and /ar/ was created by using the synthesis parameters from Holt and Lotto in the parallel branch of the Klatt (1980) synthesizer. Amplitudes for all formants other than $F3$ were set to zero. This resulted in a 250-ms harmonic complex ($f0$ equals 110 Hz) with a single frequency-varying amplitude peak. In terms of synthesis parameters, the frequency of this single formant was set at 2450 Hz for the first 100 ms for both contexts. The two contexts differed in the formant frequency trajectory over the final 150 ms. For the context modeling /al/ (referred to as *highfreq*), the formant increased linearly in frequency to 2700 Hz. For the context modeling /ar/ (*lowfreq*), the formant decreased linearly to 1600 Hz. These context sounds are not perceived as speech and certainly contain no identifiable phonemic content.

All stimuli were synthesized with 16-bit resolution at a 20-kHz sampling rate and stored on a computer disk following synthesis. Stimulus presentation was under the control of a microcomputer and Tucker Davis Technologies (TDT) hardware. Context sounds and target syllables were appended online with a 50-ms intervening silent interval. Following D/A conversion (TDT, DD1), stimuli were low-pass filtered at a 9.8-kHz cutoff frequency (TDT, FTG2), attenuated (TDT, PA4), and presented over headphones (Sennheiser HD 285) at 75 dB SPL (A).

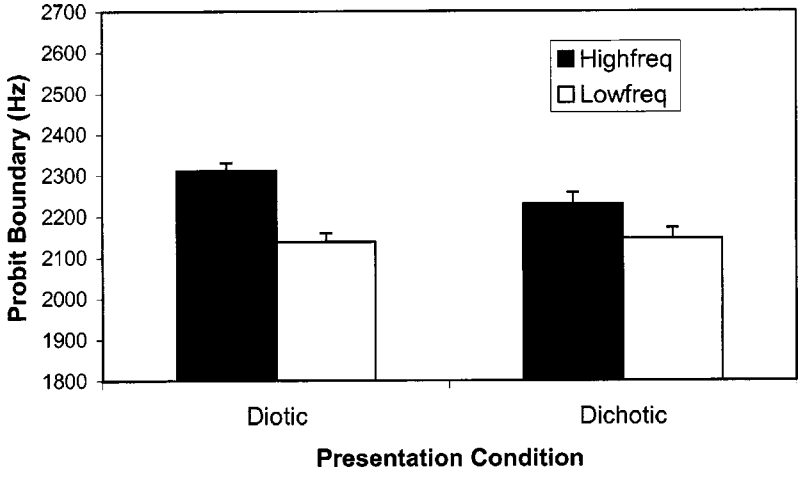

FIG. 1. Boundaries for identification of /ga/–/da/ syllables preceded by *highfreq* (dark bars) and *lowfreq* (light bars) for diotic and dichotic presentation conditions. Taller bars (higher-frequency boundaries) indicate more "ga" responses. A difference in bar height reflects an influence of preceding context on consonant identification.

TABLE I. Means and standard deviations (in parentheses) of identification boundaries as a function of context and silent interval duration from experiment 2. Differences between contexts were tested with paired-sample t-tests.

Context	25 ms	50 ms	100 ms	175 ms	275 ms	400 ms
Highfreq	2451.4 (101.8)	2403.7 (102.8)	2322.9 (124.1)	2311.4 (64.7)	2268.7 (141.7)	2296.1 (98.2)
Lowfreq	2300.4 (146.3)	2307.6 (140.5)	2232.8 (109.7)	2256.2 (87.2)	2257.7 (101.3)	2291.7 (109.2)
t-test df=16	3.64	2.74	2.92	2.33	0.36	0.14
p-value	0.0022	0.014	0.0099	0.033	0.72	0.89

3. Procedure

One to three subjects were tested concurrently in a sound-attenuated booth during a single experimental session. During each trial, listeners heard the appended stimuli (context followed by target syllable) over headphones. The listeners' task was to identify the target syllable as ''da'' or ''ga'' by pressing a labeled button on an electronic response box. Intertrial interval was approximately 3 s.

The experiment was divided into two blocks corresponding to diotic and dichotic presentation. Each subject completed both blocks and order of block presentation was counterbalanced across subjects. In the dichotic block, context and target were presented to opposite ears, with ear of context presentation randomized across trials. In the diotic block, both context and target were presented to *both* ears on each trial. In each block, listeners responded to 10 repetitions of each of the context/target combinations (2 contexts×10 target CVs×10 repetitions=200 trials per block). In all, the experiment lasted approximately 45 min.

B. Results and discussion

Previous context effect experiments (e.g., Lotto and Kluender, 1998) have used a performance criterion for inclusion of data in analyses. For the current two experiments, data were withheld from analyses for subjects who failed to correctly identify the two endpoint CVs (the best /da/ and /ga/) at least 80% of the time across conditions. In experiment 1, this led to the exclusion of data from two subjects. Identification boundaries were computed on the percentage of ''ga'' responses through probit analysis. These boundaries (in terms of $F3$ frequency of the CV series) are presented in Fig. 1. In the diotic presentation condition, identification boundaries significantly shifted from *highfreq* (2310.4 Hz) compared to *lowfreq* (2138.1 Hz) contexts [$t(21)=6.74$, $p<0.0001$]. An identification shift was also present for the dichotic presentation condition [from 2229.7 to 2145.7 Hz; $t(21)=2.24$, $p<0.05$]. A 2 (presentation condition)×2 (context) repeated measures ANOVA revealed that the effect of context was significantly greater in the diotic presentation condition [$F(1,21)=4.665$, $p<0.05$].

The results of experiment 1 are consistent with the results of experiment 1b of Holt and Lotto (2002). In the latter study, speech context effects were present for both diotic and dichotic presentation, but the effect of context was numerically smaller in the dichotic condition. A 2 (speech versus nonspeech)×2 (presentation condition)×2 (context) mixed-model ANOVA confirmed the agreement of the results from the two experiments. There were no significant interactions including the speech versus nonspeech variable ($ps>0.10$). This agreement of results suggests that similar mechanisms may underlie both speech and nonspeech context effects on phonemic identification. In both cases, it is unlikely that the identification shifts are caused solely by peripheral masking or auditory enhancement, as these mechanisms are monaural in nature. It is still possible that these peripheral mechanisms play some role in both speech and nonspeech context effects since both effects are smaller when context and target cannot interact in the periphery.

III. EXPERIMENT 2 (SILENT GAP DURATION)

A. Methods

1. Subjects

Twenty undergraduates at Washington State University participated for course credit. All were native speakers of English that reported no hearing deficits or disorders. None of the subjects participated in experiment 1.

2. Stimuli

Stimuli were identical to those used in experiment 1. Only the duration of the intervening silent interval differed. The six intervening silent intervals were 25, 50, 100, 175, 275, and 400 ms. These duration intervals are identical to those used in experiment 2b of Holt and Lotto (2002).

3. Procedure

The task for the subjects was the same as in experiment 1. Each subject participated in three blocks of 120 trials (2 contexts×6 gap durations×10 CV target stimuli). Within each block, presentation order of stimuli was randomized. The context and target stimuli were presented to both ears.

B. Results and discussion

Data from three subjects who failed to identify 80% of endpoint stimuli were excluded from further analysis. Probit boundaries for each gap duration×context condition are presented in Table I.

Planned paired-sample t-tests were used to examine the context effect at each duration of intervening silence. The spectral content of the contextual sound caused a significant shift in identification boundaries for all silent gap durations up to and including 175 ms ($ps<0.05$; see Table I). No

effect of context was present for the 275- and 400-ms gap conditions ($ps > 0.72$). Qualitatively equivalent results were obtained for tests computed on the mean percent of ''ga'' responses.

This pattern of results is quite similar to that obtained by Holt and Lotto (2002) in experiment 2b. They found an effect of speech context on CV identification out to 275 ms of intervening silence. No effect was present for a 400-ms silent gap. In both the nonspeech and speech context experiments, the size of the context effect decreases monotonically with increasing gap duration.

IV. GENERAL DISCUSSION

The pattern of results from both experiments described here matches the pattern obtained by Holt and Lotto (2002) with speech contexts. A significant context effect on CV identification remains when context is presented contralaterally to target. In both cases, the dichotic context effect is robust though smaller than for diotic presentation conditions. Effects of context also remain for substantial durations of intervening silent gaps. For speech contexts, this gap can extend to at least 275 ms. For nonspeech contexts, significant shifts were demonstrated out to 175 ms.

Fowler *et al.* (2000) propose that speech and nonspeech context effects are different in kind. However, the agreement of the current results with those of Holt and Lotto (2002) implicates similar mechanisms in both kinds of context effects. This agreement can be added to the mounting evidence for a general auditory role in speech context effects. Several previous studies have demonstrated nonspeech context effects that are equivalent in size of boundary shift to corresponding speech context effects (Lotto and Kluender, 1998; Holt *et al.*, 2000). The current studies extend these similarities across a series of presentation manipulations.

The results of these experiments support the contention of Lotto and Kluender (1998; Lotto *et al.*, 1997) that general mechanisms of the auditory system are at least partially responsible for the kinds of speech context effects examined here. The result of these general mechanisms is the perceptual emphasis of energy in frequency regions that are less represented in context sounds. That is, changes in the pattern of spectral energy are enhanced. The behavioral input–output function can be described as *spectral contrast* and it appears to be a general property of auditory systems. Lotto *et al.* (1997) demonstrated that birds (Japanese quail, *Coturnix japonica*) trained to respond to /da/ and /ga/ stimuli also show contrastive response shifts with /al/ and /ar/ contexts. Lotto and Kluender's description of the pattern of contrastive output does not implicate any particular mechanism. However, the results of the current set of experiments provide evidence against some proposed mechanisms.

Given the monaural nature of peripheral masking and auditory enhancement, it is unlikely that either of these mechanisms is solely responsible for context effects. The fact that dichotic context effects were smaller suggests that it is possible that peripheral mechanisms play *some* role. However, a complete explanation will require a description of more central processes that take input from both ears. The relative temporal robustness of the context effects described in experiment 2 is also consistent with a central mechanism. In general, as one observes effects of interactions at more central levels of the auditory system, there is a longer temporal window over which auditory events interact and influence one another (Popper and Fay, 1992). These results are in agreement with neurophysiological investigations of speech context effects that found little evidence for contrast at the auditory nerve (Holt and Rhode, 2000).

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. **26**, 877–888.

Holt, L. L., and Lotto, A. J. (**2002**). ''Behavioral examinations of the neural mechanisms 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.

Holt, L. L., and Rhode, W. S. (**2000**). ''Examining context-dependent speech perception in the chinchilla cochlear nucleus.'' Paper presented at the 2000 Midwinter Meeting of Association for Research in Otolaryngology, St. Petersburg Beach, FL.

Klatt, D. H. (**1980**). ''Software for a cascade/parallel formant synthesizer,'' J. Acoust. Soc. Am. **67**, 971–990.

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 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.

Popper, A. N., and Fay, R. R. (**1992**). *The Mammalian Auditory Pathway: Neurophysiology* (Springer, New York).

Summerfield, Q., and Assmann, P. F. (**1989**). ''Auditory enhancement and the perception of concurrent vowels,'' Percept. Psychophys. **45**, 529–536.

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

Viemeister, N. F. (**1980**). ''Adaptation of masking,'' in *Psychophysical, Physiological, and Behavioral Studies in Hearing,* edited by G. van den Brink and F. S. Bilsen (Delft University Press, Delft, The Netherlands), pp. 190–199.

Viemeister, N. F., and Bacon, S. P. (**1982**). ''Forward masking by enhanced components in harmonic complexes,'' J. Acoust. Soc. Am. **71**, 1502–1507.