Cognition 108 (2008) 710–718

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

Talker adaptation in speech perception: Adjusting the signal or the representations?

Delphine Dahan [a,*], Sarah J. Drucker [a], Rebecca A. Scarborough [b]

[a] Psychology Department, University of Pennsylvania, PA, USA
[b] Linguistics Department, University of Colorado at Boulder, CO, USA

ARTICLE INFO

Article history:
Received 7 August 2007
Revised 10 June 2008
Accepted 13 June 2008

Keywords:
Speech perception
Speaker adaptation
Eye movements

ABSTRACT

Past research has established that listeners can accommodate a wide range of talkers in understanding language. How this adjustment operates, however, is a matter of debate. Here, listeners were exposed to spoken words from a speaker of an American English dialect in which the vowel /æ/ is raised before /g/, but not before /k/. Results from two experiments showed that listeners' identification of /k/-final words like *back* (which are unaffected by the dialect) was facilitated by prior exposure to their dialect-affected /g/-final counterparts, e.g., *bag*. This facilitation occurred because the competition between interpretations, e.g., *bag* or *back*, while hearing the initial portion of the input [bæ], was mitigated by the reduced probability for the input to correspond to *bag* as produced by this talker. Thus, adaptation to an accent is not just a matter of adjusting the speech signal as it is being heard; adaptation involves dynamic adjustment of the representations stored in the lexicon, according to the characteristics of the speaker or the context.

1. Introduction

Speech is a complex and highly ambiguous signal. A major source of ambiguity is the impact that contextual factors, such as the speaker's characteristics, have on the acoustic realization of words. Nonetheless, human listeners, even very young ones, have the capacity to recognize familiar words from the speech of any speaker of their language, with no prior exposure to this speaker (Hallé & de Boysson-Bardies, 1994; Swingley, 2005). Furthermore, a large body of research has documented listeners' remarkable plasticity in accommodating speaker-specific, sometimes peculiar, pronunciations (e.g., Clarke & Garrett, 2004; Eisner & McQueen, 2005; Kraljic & Samuel, 2006; Maye, Aslin, & Tanenhaus, 2008; Norris, McQueen, & Cutler, 2003). For the most part, this research has examined *what* people learn that may allow them to generalize their experience with specific tokens to new instances. An aspect of perceptual learning that has received relatively less attention is *how* talker adaptation is achieved.

The traditional approach to cross-speaker variability assumes that word recognition operates on speaker-normalized input. A number of normalization algorithms have been proposed, in which information extracted from the speech input is transformed either to neutralize the influence of speaker variability, in effect treating it as noise, or to model talker-specific variation and factor out its influence. The modeling of the effects of vocal-tract length on vowels' formant frequencies is an example of the latter (see Johnson, 2005, and references therein). On *normalization* accounts, then, the acoustic signal is warped to match lexical knowledge that is stored in an abstract, speaker-independent, and immutable form. Talker adaptation, according to this view, consists of learning a new way to normalize the speech signal.

Another mechanism to talker adaptation, however, assumes that the representations that listeners evaluate when processing speech are dynamically altered or assembled to reflect the current context, including the identity of

* Corresponding author. Address: Psychology Department, Room 302C, 3401 Walnut Street, Philadelphia, PA 19104-6228, USA. Tel.: +1 215 898 0326; fax: +1 215 573 9247.
E-mail address: dahan@psych.upenn.edu (D. Dahan).

0010-0277/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.cognition.2008.06.003

the talker. We will refer to this as the *representational* account. The form that adaptation takes may vary depending on the nature of word-form representations one assumes. If words are viewed as collections of context-marked experienced instances (Goldinger, 1998; Johnson, 1997), context- or talker-adaptation is a basic property of those theories, as memory traces associated with a particular context or talker are primarily recruited when processing speech from the same context or talker. If, on the other hand, word representations are viewed as having abstracted away from direct experiences, either by representing the central tendency of those instances or by capturing the defining properties of word forms (representations that most phonological theories assume), talker adaptation consists of temporarily altering those representations or the representations of their components.

Although the normalization and representational accounts have been acknowledged in the literature, the evidence that unambiguously supports one over the other has been limited. For instance, evidence that listeners' familiarity with a talker affects processing of that talker's speech (e.g., Goldinger, 1996; Ladefoged & Broadbent, 1957; Nygaard & Pisoni, 1998) is compatible with both the representational and normalization views of speech processing because normalization algorithms themselves may adjust. Likewise, evidence that listeners learn from their experience with a peculiar speech sound and can generalize it to new instances of the same sound (e.g., Kraljic & Samuel, 2006; McQueen, Cutler, & Norris, 2006) does not in itself demonstrate that listeners alter their representation of speech-sound categories; the finding is equally compatible with an account where the normalization algorithm has been altered to warp the speech signal to fit a context-general representation of sound categories.

Thus, the two possible mechanisms underlying talker adaptation in speech processing have yet to be convincingly distinguished empirically. Although our understanding of the process by which listeners adapt to a foreign accent or artificially distorted speech has benefited considerably from the growing number of studies documenting this plasticity, this work has not directly contrasted these two mechanisms. The current study is an attempt to do so.

In some American English dialects, the vowel /æ/ before /g/ (but not before /k/) raises to a vowel approaching [ɛ], as in *beg*. This phenomenon, in effect, reduces the phonetic overlap between word pairs like *bag* and *back*. Phonetic overlap in words is functionally significant because listeners begin generating hypotheses about the word they are hearing from the moment the word begins (Marslen-Wilson, 1973). All else being equal, the beginning of *back*, [bæ], is equally compatible with the word *back* or *bag* (among others). Here, we asked if listeners' prior exposure to a speaker's raised vowel in words like *bag* affects the subsequent processing of their unaffected counterpart words, e.g., *back*.[1] Such an effect would reflect the reduced probability for the input [bæ] to correspond to the initial portion of the word *bag* produced by this speaker (in Bayesian terms, the likelihood of the input [bæ], given the hypothesis *bag*).

Importantly, this effect would reflect the influence of listeners' speaker-specific mental representations on the assessment of standard unaccented spoken input. Thus, such a finding could not be explained by a transformation of the spoken input to fit speaker-independent representations – there is no transformation to undertake, because the heard form already matches the listener's lexicon. Rather, it would suggest that representations are dynamically adjusted to the current speaker or context. Thus, the question at stake here is not whether listeners can learn to adapt to the speaker's non-standard pronunciation of *bag*; previous work on speaker adaptation and perceptual learning has convincingly demonstrated that listeners can. Rather, we asked whether we can observe the consequences of this learning when the speech signal does not *require* adjustment, i.e., when the pronunciation of the spoken word is the standard one. This provides an informative test of the mechanism of perceptual adaptation.

Participants were presented with a series of trials on which they saw four written words on a computer screen (e.g., *bag*, *back*, *wig*, *wick*) and heard one of the words spoken. Their task was to indicate which word they heard by clicking on it with the computer mouse. Participants' eye movements to the written words were recorded as the target name was heard and until participants clicked on one of the words. Language-guided eye gaze was used as the dependent measure because it reflects listeners' ongoing, moment-by-moment interpretation of the speech signal, without imposing a distracting secondary task (Dahan, Magnuson, & Tanenhaus, 2001; Tanenhaus, 2007). As listeners begin to hear a word like *back*, their interpretation of [bæ...] in the signal as referring to *back* or *bag* can be inferred from their fixation behavior.

Listeners' ability to adapt their internal representations to their past experience with a given speaker was examined in two experiments. Experiment 1 contrasted two groups of listeners: Those exposed to /g/-final words with a standard vowel, and those exposed to such words with a raised vowel. Both groups were then tested on the same /k/-final words like *back*. Experiment 2 adopted a within-subject design and compared listeners' recognition of /k/-final words before and after exposure to /g/-final words with a raised vowel. Experiment 1 used auditory stimuli where vowel duration, an acoustic correlate of the identity of the subsequent consonant, had been neutralized to maximize the ambiguity between the /k/ and /g/ interpretations. When confronted with potentially ambiguous tokens of the *back*-like words, we reasoned, listeners may be especially likely to recruit all potential cues to resolve the ambiguity. If listeners can adapt their representation to their past experience, those who were exposed to /g/-final words with a raised vowel would be less likely to misinterpret the *back*-like words (with a standard vowel) as their *bag*-like counterparts than listeners who were exposed to /g/-final words with the standard vowel. To assess that this recruitment was not restricted to the processing of ambiguous stimuli, Experiment 2 used only tokens with their original vowel durations.

[1] Technically, all words produced by a speaker are affected by his/her dialect. Our use of the term "unaffected" throughout the paper should be taken to reflect the fact that these words' pronunciation matches that found in standard American English.

2. Experiment 1

2.1. Methods

2.1.1. Participants

Fifty students at the University of Pennsylvania participated. All were native speakers of American English and received course credit. Participants were randomly assigned to the raised-vowel or standard-vowel condition, as described below.

2.1.2. Stimuli

Eleven monosyllabic word pairs containing the vowel /æ/ and ending with either /k/ or /g/ were selected (e.g., *back* and *bag*). Although the frequencies of the /k/- and /g/-final words could not be matched, all items were selected to be highly familiar to college students (Nusbaum, Pisoni, & Davis, 1984). An additional 11 word pairs served as filler items; these words also ended in /g/ or /k/ but contained vowels other than /æ/ (e.g., *wig* and *wick*). Across pairs, the /k/ and /g/ filler items were roughly matched for frequency (see Appendix for the list of items).

All words were read aloud by a female native speaker (the third author) of the American English dialect in which *bag*-like words are spoken with a raised vowel and *back*-like words, with the standard [æ]. Stimuli were recorded directly to a computer and edited using the Praat speech-analysis software (Boersma & Weenink, 2005). *Bag*-like words with the standard [æ] vowel were created by replacing the [k] consonant from the *back*-like word with the [g] consonant from its raised *bag*-like counterpart. The splicing point corresponded to the end of the vowel and the beginning of the closure. Because of the possible artifacts caused by the splicing manipulation itself, and to avoid confounding the use of spliced stimuli with the group participants were assigned to, raised *bag*-like stimuli were also manipulated by splicing two different tokens of the same word together. The same procedure was also applied to the standard *back*-like words and all of the filler items. The final stimulus set thus contained two versions of the *bag*-like words, one with the standard vowel (cross-spliced) and the other, the raised vowel (identity-spliced), as well as a single version of the *back*-like words and all filler words (identity-spliced).

Because vowels are longer before /g/ than before /k/ in English (House, 1961), our cross-splicing manipulation on the *bag*-like words required the adjustment of vowel duration to achieve natural-sounding tokens. Preliminary attempts revealed that the expansion of the short vowel [æ] from the original *back*-like word (on average, 246-ms long) to match, within the pair, the long raised vowel's original duration from the *bag*-like word (on average, 354-ms long) resulted in unnatural-sounding tokens. As a compromise, we lengthened the [æ] vowels to reach an intermediate value between the originally short and long durations. The equivalent manipulation was applied to the raised vowels of the original *bag*-like words, shortening them to match the intermediate value. Finally, we lengthened the vowels in the identity-spliced *back*-like words, to match the duration of their *bag*-like counterparts. This in effect neutralized vowel duration as a cue to the identity of the following consonant in order to amplify the ambiguity between the /g/-final (e.g., *bag*) and /k/-final (e.g., *back*) interpretations of the *back*-like spoken words. Vowel-duration neutralization was also performed on the filler items.

To lengthen vowels, an individual period in the acoustic waveform was selected and duplicated. This was repeated at evenly spaced intervals throughout the vowel until the desired length was attained. To shorten vowels, periods were removed using the same procedure. Vowel duration for the final stimuli was 298 ms on average for the critical items and 221 ms for the fillers; within each pair, the vowels were matched in duration.

2.1.3. Procedure

Participants were instructed that on each trial they would hear one of the four words displayed on the screen; their task was to click on it with the computer mouse and move it near the geometric shape adjacent to it. At the start of every trial, a 5×5 grid with four printed words, four geometric shapes, and a central fixation cross appeared on the screen. The words were positioned to form the corners of an imaginary square on the grid, with a geometric shape flanking each of them. After 750 ms, the auditory target word was presented through headphones. Once the move was completed, the experimenter pressed a button to proceed to the next trial. Eye movements were monitored with a head-mounted eye-tracker (Eyelink II, SR Research), sampling at 250 Hz.

Each critical pair always appeared with the same filler pair. Each display was presented 10 times, varying whether the target word was an item from the critical pair (2 times per item) or from the filler pair (3 times per item). The resulting 110 trials were organized into four blocks. In block 1, participants heard each of the *bag*-like items from the 11 critical pairs once (with the standard vowel for half of the subjects, and the raised vowel for the other half). In block 2, each of the 11 *back*-like items was mentioned once. Blocks 3 and 4 replicated blocks 1 and 2, respectively. Filler items were interspersed so that subjects heard both /g/-final and /k/-final targets in each block. Four trial lists were created in which the trial order within each block and the position of the printed words in the grid on each trial were randomized. Within each list, the version of the *bag*-like words was consistent (i.e., with a standard or raised vowel). Participants were randomly assigned to each list and version.

Each trial was coded in terms of the correctness of the word selected as the target (as evidenced by the mouse click) and the fixations that participants made as they heard and processed the spoken word (see Dahan & Gaskell, 2007, for details on defining and categorizing fixations). Analyses were limited to fixations taking place from the onset of the spoken word's vowel until a fixation to the target word was made (which was taken as evidence of listeners' selection of the target word as the referent), provided that this fixation ended no earlier than 500 ms after the spoken-word vowel onset. This criterion was adopted to minimize the risk of terminating the coding of a trial's fixations prematurely.

2.2. Results

Fifty-one trials (i.e., 2.3% of the critical trials) were excluded from subsequent analyses because of technical failure or track loss (16 trials) or because participants failed to make a fixation on the target word by the end of the trial (35 trials).

Analysis of subjects' errors on the trials from blocks 2 and 4, in which both groups responded to the same *back*-like words, showed that participants in the standard-vowel condition were substantially more likely to misidentify *back*-like words as their *bag*-like counterparts than participants in the raised-vowel condition (17.9% vs. 7.3%). An analysis of variance (ANOVA), performed on the arcsine-transformed error rates computed for each participant and on each block revealed a significant effect of group ($F(1,48) = 5.5$, $p < .05$, $\eta^2_G = 0.09$), with no main effect of block and no interaction.[2] Importantly, the two groups did not significantly differ on the identification of the *bag*-like words, with an error rate of 0.9% for the standard group and of 0.2% for the raised group.[3]

Thus, the ambiguity that resulted from neutralizing vowel duration as a cue to the identity of the subsequent consonant was significantly mitigated by listeners' past exposure with the speaker's pronunciation of the *bag*-like words: Participants were less likely to confuse *back*-like words for their *bag*-like counterparts when participants had heard the speaker pronounce the latter with a different vowel. This effect provides strong support to the hypothesis that listeners evaluate the spoken input with respect to dynamically adjusted, context-specific representations. Exposure to the talker's speech affected listeners' interpretation of sounds that were not accented, showing that adaptation was not a kind of "correction" of the talker's speech, but rather a change in the phonetic properties of the representations associated with the *bag*-like words.

Analyses of the trials on which participants correctly identified the *back*-like words provided additional support for this hypothesis. Fixation proportions were computed by defining 50-ms time bins and counting the number of trials where a fixation to each of five regions (either one of the four words or elsewhere on the screen) occurred during each time bin, for each participant and for each block. Fig. 1 displays the proportion of fixations to the target word (e.g., *back*) and its onset-overlapping competitor (e.g., *bag*) over time, from the onset of the vowel of the *back*-like spoken words, averaged over blocks 2 and 4. From about 600 ms on, participants in the standard-vowel group fixated the competitor word for a more extended period of time than participants in the raised-vowel group. This shows that the former group wrongfully entertained the possibility of hearing the word *bag* more often and for longer than the latter group, and consequently, were slower at identifying (and thus, fixating on) the target word.

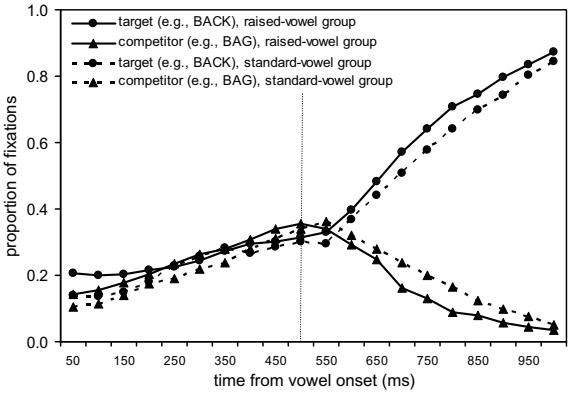

Fig. 1. Experiment 1: Proportion of trials with a fixation to the target word (circle symbols, e.g., BACK) and to the competitor word (triangle symbols, e.g., BAG) over time, from the onset of the spoken-word vowel until 1000 ms, as a function of the vowel in *bag*-like words participants were previously exposed to (i.e., raised vowel, solid line, or standard vowel, dashed line). The vertical line marks the onset of the 500–1000 ms analysis window.

To statistically confirm this effect, we computed the average proportion of fixations to the target and competitor over two separate time windows, one extending from 0 to 500 ms, and the other, from 500 to 1000 ms. The 500-ms boundary corresponds to the point at which the impact of having processed the vowel should be most observable given the duration of the vowel (300 ms on average) and a well-established 200-ms estimated delay in programming and launching a saccade (Hallet, 1986). Thus, most robust effects of vowel quality should be observed on the 500–1000 ms window. We derived a measure that encompasses the effect on both the target and competitor fixations by subtracting the competitor proportion from the target proportion. An ANOVA performed on the fixation–proportion differences computed over the 0–500-ms window revealed no main effect and no interaction. However, the ANOVA performed on fixation-proportion differences computed over the 500–1000 ms window revealed a main effect of group ($F(1,48) = 4.7$, $p < .05$, $\eta^2_G = 0.06$), no effect of block and no interaction.[4] Thus, the experience of hearing a speaker raise the vowel in the context of /g/ affected listener's interpretations of words that were themselves unaffected by the dialect, i.e. the *back*-like words.

To ensure that the group effect on gaze behavior reflected participants' experience with raised vs. standard vowels in *bag*-like words, and not general differences in gaze behavior between the two groups (a common concern with between-subject designs), we compared participants' performance on the critical /k/-final trials to that on the filler trials. If the eye-gaze difference observed on the critical /k/-final word trials between the two groups specifically resulted from exposure to their /g/-final counterparts, we

[2] Following Bakeman (2005), we report the effect-size statistics η^2_G to provide comparability across between- and within-subjects designs.

[3] The fact that our vowel-duration neutralization affected the identification of *back*-like words substantially more than the identification of *bag*-like words may reflect the fact that the duration distribution of vowels perceived as preceding a voiced consonant like [g] can have a wider range of durations than vowels perceived as preceding a voiceless consonant like /k/ (Raphael, 1972).

[4] Here and throughout the paper, analyses conducted on the arcsine-transformed fixation proportions yielded identical results, unless otherwise specified in the text.

reasoned, eye gaze should not differ between the two groups on filler trials, which were identical. For each participant, we assessed performance on the 66 filler trials by computing the difference in fixation proportions to target and competitor over the 500–1000 ms window. A two-way ANOVA (with group as a between-subject factor and trial type – critical or filler – as a within-subject factor) revealed a main effect of group (i.e., larger target-competitor difference for the group exposed to raised *bag*-like words than for the group exposed to standard *bag*-like words, $F(1,48) = 11.5$, $p < .001$, $\eta^2_G = 0.15$), and a main effect of trial type (larger target-competitor difference on fillers than on critical trials, which indicates more modest competition from the competitor over the target on filler than critical trials, $F(1,48) = 115.4$, $p < .001$, $\eta^2_G = 0.4$). This weaker competition effect on filler than critical trials may be attributed to the shorter vowel duration for the filler items (221 ms) than for the critical items (298 ms). Shorter vowel duration implies that disambiguation of the target over the competitor can take place earlier in time, yielding a greater target-competitor difference in fixation proportions over the 500–1000 ms window. Critically here, the analysis also revealed a significant interaction between group and trial type ($F(1,48) = 12.8$, $p < .001$, $\eta^2_G = 0.07$). Separate one-way ANOVAs confirmed this, revealing a significant effect of group on data from the critical trials ($F(1,48) = 16.0$, $p < .0001$, $\eta^2_G = 0.25$) but not on data from the filler trials ($F(1,48) = 1.7$, $p > .20$). Thus, the group difference in the processing of *back*-like words cannot be attributed to overall differences in gaze behavior between the two groups.

Although not central to the focus of the study, eye gaze on trials where people heard *bag*-like words with standard or raised vowels were also analyzed. Fig. 2 displays the proportion of fixations to the target (e.g., *bag*) and competitor (e.g., *back*) for each participant group, averaged over blocks 1 and 3. The figure indicates reduced competition and faster target identification when participants heard the raised-vowel *bag*-like words than when they heard the standard-vowel *bag*-words. The group difference that resulted from hearing different pronunciations of the *bag*-like words arose early in time, somewhat earlier than the differences observed on the /k/-final trials where both groups heard the same *back*-like words and where differences between the groups can only be attributed to their past exposure with the talker's utterances. Although the apparent difference in timing between the two types of trials is intriguing, the group difference on the *bag*-like words was not statistically confirmed: The ANOVA performed over the 0–500 ms time window yielded a main effect of group when the differences between fixation proportions to target and to competitor were expressed in untransformed proportions ($F(1,48) = 4.9$, $p < .05$, $\eta^2_G = 0.04$), but not when computed over arcsine-transformed proportions. Because the arcsine transformation protects from possible spurious results that ANOVAs conducted on untransformed proportions may yield, analyses based on the former are more valid. Analyses performed on fixation proportions computed over the 500–1000 ms window, on the other hand, revealed a robust main effect of group ($F(1,48) = 8.2$, $p < .01$, $\eta^2_G = 0.1$), no effect of block, and no interaction. This result confirms that the raising of the vowel reduces the overlap between the *bag*-like words and their *back*-like counterparts. Importantly, the fact that the data did not differ across blocks indicates that listeners learned to associate a raised vowel with the pronunciation of *bag*-like words very rapidly, based on the first few instances.

Experiment 1 provides strong evidence that listeners who were exposed to the raised pronunciations of *bag*-like words have altered their representation of how they expect these words to sound and bring these representations to bear when processing unaltered words from the same talker.

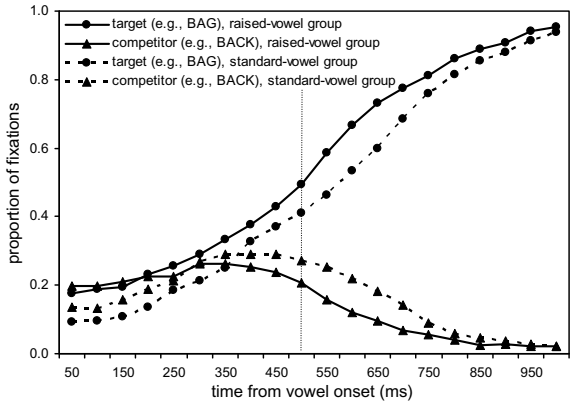

Fig. 2. Experiment 1: Proportion of trials with a fixation to the target word (circle symbols, e.g., BAG) and to the competitor word (triangle symbols, e.g., BACK) over time, from the onset of the spoken-word vowel until 1000 ms, as a function of the vowel quality in the spoken words (i.e., raised vowel, solid line, or standard vowel, dashed line). The vertical line marks the onset of the 500–1000 ms analysis window.

3. Experiment 2

Experiment 2 aimed to replicate Experiment 1 using a within-subjects design. A single group of participants was tested on the identification of *back*-like words before and after exposure to the speaker's dialect-affected pronunciation of *bag*-like words. Before this exposure, we reasoned, listeners should assume a standard pronunciation for the *bag*-like words, and this should result in substantial, albeit temporary, ambiguity during the identification of *back*-like words. After exposure to the raised pronunciation of *bag*-like words, however, this ambiguity should lessen if participants evaluate the speech signal with respect to representations that reflect their context-specific experience. To establish that such a change could not be accounted for by increased practice with the task or the items, the same repeated presentation was applied to filler item pairs, for which the dialectal vowel raising did not apply. For those fillers items, only practice with the task or items may affect word recognition. We thus predicted facilitation in the identification of *back*-like words after exposure to raised *bag*-like words, above and beyond any potential impact of practice.

We also sought to replicate Experiment 1's results with stimuli where vowel duration, a cue to the identity of the final consonant, was left intact. Unmanipulated tokens of the *back*-like words should be recognized as such and error rates should be minimal. We asked if, under these conditions, eye movements to target and competitor words taking place during the recognition of the *back*-like words reflect evidence that listeners recruit context-specific representations.

3.1. Methods

3.1.1. Participants

Forty-two participants were recruited using the same procedure as in Experiment 1.

3.1.2. Stimuli and procedure

The stimuli consisted of the same tokens as those used in Experiment 1 (11 critical word pairs and 11 filler pairs) but devoid of any splicing or vowel-duration manipulation.[5] Vowel duration for the /k/ and /g/ critical words was on average 252 and 354 ms, respectively; vowel duration for the /k/ and /g/ filler words was on average 191 and 256 ms, respectively.

Throughout the experimental session, each four-item display (a critical pair associated with a filler pair) was presented 8 times, with each word serving as the target twice. Within each pair, the order with which each word was the target was fixed, starting with the /k/ item and then alternating with its /g/ counterpart. Because the presentations of the filler and critical items were manipulated in the same way, trials were not blocked but interspersed, carefully controlling for the lag between the first and second presentation of the /k/ item of each pair. Eight such trial orders were created, and across these orders, the lag between the first and second presentation of the /k/ item for both the critical and filler pairs was 40 trials. Participants were randomly assigned to each random order.

The eye-tracking, stimulus-presentation, and coding procedures were identical to those in Experiment 1. Analyses were restricted to the 44 trials where the /k/ item of each 11 critical and filler pairs was the target.

3.2. Results

Fifty-nine trials (3.2% of the data) were excluded from all analyses because of technical failure or track loss (4 trials) or because participants failed to fixate on the referent word by the end of the trial (55 trials). On 29 of the remaining trials (1.6%), participants erroneously identified the /k/ words as their /g/ counterparts. In comparison to what was observed in Experiment 1, this low error rate indicates that the lexical ambiguity attributed to manipulating the vowel duration was largely mitigated. Nonetheless, errors were not equally distributed across conditions: While the number of such errors remained low on first and second presentations of the /k/ word on the filler pairs (4 and 2, respectively), the number of such errors was greater and dropped significantly between the first and second presentation of the /k/ words on the critical pairs (20 vs. 4, $p < .001$ based on a binomial test, with 14 vs. 2 out of 42 participants making an error on the first vs. second presentation). Although it is unclear why participants, on their first presentation, misidentified the *back*-like words for their *bag*-like counterparts more often than they confused filler words (4.3% vs. 0.9%), it is remarkable to note that this tendency disappeared on the second presentation of the *back*-like words. This effect could be attributed to participants' experience with the talker's pronunciation of those words and/or, more interestingly, experience with the talker's pronunciation of their /g/ counterparts with a non-standard, raised vowel. (Only one error was observed over all critical and filler /g/-word trials.) All error trials were removed from subsequent analyses.

Fig. 3 presents the proportion of fixations to the target and competitor words upon hearing the *back*-like words before and after exposure to their raised *bag*-like counterparts. As is apparent in the graph, from about 600 ms onward, participants oriented their gaze toward the target faster and fixated on the competitor less on their second presentation of the test words, thus after they had heard the speaker say the *bag*-like words, than on their first presentation. This effect was especially strong on the target fixations (solid line over the dashed line) but was also present, though more modestly, in competitor fixations (solid line under the dashed line).

To test the significance of this effect and similarly to our approach in Experiment 1, we computed, on /k/ critical and filler trials, the difference between target and competitor fixation proportions for each time bin, and averaged them over a 500–1000 ms window. The two-way (trial type × presentation) ANOVA revealed only a main effect of presentation (with a larger target-competitor difference on the second presentation of the /k/ trial than on its first presentation ($F(1,41) = 5.2$, $p < .05$, $\eta^2_G = 0.03$) but no significant interaction between the trial type and presentation. However, one-way ANOVAs conducted on the

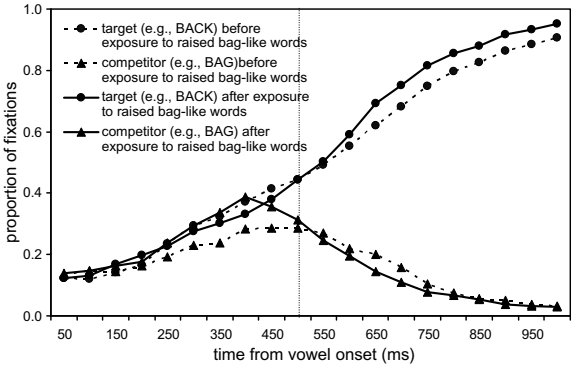

Fig. 3. Experiment 2: Proportion of trials with a fixation to the target word (circle symbols, e.g., BACK) and to the competitor word (triangle symbols, e.g., BAG) over time, from the onset of the spoken-word vowel until 1000 ms, before (dashed lines) and after (solid lines) exposure to raised *bag*-like words. The vertical line marks the onset of the 500–1000 ms analysis window.

[5] In the few cases where the need for creating natural-sounding spliced stimuli in Experiment 1 had precluded the selection of the best token from our original recording, that best token was selected here.

critical trials and on the filler trials separately showed a reliable effect of presentation on critical trials ($F(1,41) = 4.3$, $p < .05$, $\eta^2_G = 0.02$) but not on filler trials ($F(1,41) = 2.1$, $p > .10$). Taken together, these analyses indicate that, while the ability to distinguish the target from its competitor improved between the first and second presentation of all /k/ trials, the improvement was statistically robust only for the critical trials. This brings support to the hypothesis that people specifically benefited from exposure to raised *bag*-like words in their subsequent ability to distinguish the *back*-like words from their *bag*-like competitors.

Also observable in Fig. 3 is an initial tendency for participants to fixate on the competitor more on their second exposure to the test words than on their first. Although analyses conducted on the difference between target and competitor fixation proportions and averaged over the 0–500 ms window revealed no significant effect of presentation, this pattern is intriguing enough to warrant some discussion. A close examination of the fixation proportions reveal that on their first exposure to the test words (dashed lines), participants favored *back*-like interpretations (the eventual targets, shown as dashed circles) over *bag*-like interpretations (the competitors, shown as dashed triangles) as early as 200 ms after the onset of the vowel. Because this preference takes place too early to reflect uptake of disambiguating information in the signal, it most likely reflects a bias toward *back*-like words, owing to the fact that the *back*-like words are more common in English than the *bag*-like words (see Appendix). Frequency effects of this sort have a well-established influence on early fixations (e.g., Dahan et al., 2001). Over the course of the experiment, as participants heard both the *bag*- and *back*-like words, the frequency biases were attenuated, as reflected by roughly equivalent target and competitor fixation proportions early on in the later trials (solid line circles and triangles). In support of this account, the bias to fixate the /k/ word on the first trial involving each critical pair was also evident in the performance of Experiment 1's participants in the standard-vowel group, block 1. However, it had no bearing on our main analyses, which focused on the recognition of the *back*-like words in blocks 2 and 4.

Thus, Experiment 2 replicated the primary result of Experiment 1 within subjects and with stimuli that were only transiently ambiguous: Exposure to a talker whose /æ/ vowel is altered in the context of /g/ affected interpretation of words not affected by the dialect.

4. General discussion

Past research has established that listeners can adapt to the characteristics of a talker's speech. For example, listeners can learn to interpret an ambiguous sound halfway between [s] and [f] as an instance of the sound /f/ (e.g., Norris et al., 2003). In previous research, this has often been interpreted as evidence that listeners have modified their representation of the phonetic category /f/ to include the ambiguous sound. However, this behavior can also be interpreted as evidence that listeners adjust the mapping process linking heard speech to representations of categories that are themselves abstract and immutable, i.e., they learn to "correct" the ambiguous sound. These two accounts cannot be distinguished on the basis of listeners' processing of the altered sound. The innovation of the present study is to assess listeners' adaptation on their interpretation of speech that was unaffected by the speaker-specific attribute. Experiments 1 and 2 demonstrated that listeners form context-contingent representations and base their evaluation of a spoken word from the same talker on these representations, even when the spoken-word's pronunciation is unaffected by the talker's dialect idiosyncrasy. Furthermore, Experiment 2 showed that this recruitment occurs even when the spoken words were unambiguous overall. Much of the empirical work demonstrating people's perceptual plasticity has relied on globally ambiguous or bi-stable stimuli (Norris et al., 2003; Haijang, Saunders, Stone, & Backus, 2006). Our results bolster the claim that listeners routinely adapt to the speech of their interlocutor and use adjusted representations to more effectively evaluate speech.

This study's central contribution, thus, is to show that this adaptation cannot be confined to learning a novel way of interpreting non-standard or peculiar pronunciation of speech sounds or words. Listeners' adjustment was demonstrated on the processing of words unaffected by the dialect. The input-normalization approach cannot account for this finding because a normalization algorithm computed to interpret the raised vowel as an instance of the underlying category /æ/ would not be implicated in the processing of a word with the standard vowel. The representational view, on the other hand, offers a natural explanation for the current results: Past experiences with a talker's utterances may result in changes in the way listeners expect specific words (or their components) to sound. In the present case, for instance, the word *bag* has ceased to be as compatible with the spoken input [bæk] as it was before exposure to that talker's raised pronunciation of *bag*. This knowledge is brought to bear on the processing of speech from the same talker (or, more generally, from the same context), independently of whether or not the speech itself diverges from the standard pronunciation.

A number of important questions arise from the current finding. In particular, is listeners' adjustment limited to the words they were exposed to and/or to the current talker, or can their experience generalize to other words and/or to other talkers? On the one hand, listeners may be initially conservative, i.e., treating their experience as reflecting the idiosyncratic pronunciation of specific words from a specific individual; subsequently, after enough evidence of the generality of the phenomenon has accumulated (in the present case, perhaps after exposure to several different words containing a raised version of the vowel /æ/ before /g/), listeners may begin to generalize beyond their direct experience. Alternatively, listeners may initially assume that their experience exemplifies something very general about the talker (or even a speech community) and display broad generalization, possibly revising this hypothesis after counter-evidence has been encountered. For instance, listeners may immediately assume that the raised pronunciation of the vowel /æ/ applies to all pho-

netic contexts until they have had direct experience with the talker's pronunciation of non-altered /æ/ in /k/ contexts. Although these issues were not the present study's focus, some aspects of the results reported here speak to these questions. First, the fact that performance on the recognition of the *bag*-like words from Experiment 1 did not change much between their first and second presentations (i.e., blocks 1 vs. 3) suggests that, based on just a few trials, listeners very rapidly learned to associate the raised vowel with *bag*-like interpretations and were able generalize this association to the subsequent /-æg/ words within the same trial block. Second, the *back*-like words acted as weaker competitors when the *bag*-like spoken words contained a raised vowel than when they contained a standard vowel (see Fig. 2). Because this pattern was already present in block 1, thus before any instances of *back*-like words were heard, this indicates that listeners who heard *bag*-like words with a raised vowel did not assume that all instances of the vowel /æ/ would be raised by this talker – if they had, *back*-like words would have become equally strong competitors as they were for the listeners who heard *bag*-like words with a standard vowel. Taken together, these aspects of the data suggest that, with a fairly short exposure to a talker's dialectal characteristics, listeners display some, albeit limited, generalization. Although more work is necessary before any firm conclusion can be established, work on talker adaptation is increasingly revealing how variable the conditions for generalization can be, as the sample that listeners have of the talker's speech changes.

Another question that arises from the present finding is the degree to which prior exposure with the talker's dialect affects listeners' adaptation to the talker. In the present study, we chose not to select or screen participants as a function of their possible experience with the dialect for practical reasons: Asking participants to report the places where they have lived may not be a reliable indicator because the occurrence of /æ/ raising in /g/ contexts is not restricted to well-delimited areas of the United States. It has been described as a northern Midwestern and West feature (Labov, Ash, & Boberg, 2005; Zeller, 1997), but can be found in other areas, as exemplified by the current talker (the third author) who is from Kansas. An alternative to the geographic questionnaire, i.e., a recording of participants' pronunciation of the dialect-affected words, was also deemed inadequate as it may mischaracterize participants' experience with the dialect as *listeners*. Other dialectal variations, especially those whose geographic spread is well documented, may better lend themselves to the investigation of the role of prior experience with a dialect on the ability to display adaptation to it in an experimental context.

The current study has shown that adapting to a talker involves learning from the talker's specific pronunciations, incorporating this learning into representations, and later relying on this knowledge when evaluating novel utterances from the same speaker. Because evidence of adaptation was observed on the processing of words that were unaffected by the characteristics of the talker, this study provides perhaps the first definite evidence that listeners can adjust their representations dynamically.

Acknowledgments

This work was supported by the National Science Foundation under Grant No. 0433567 and the National Institutes of Health (1 R01 HD 049742-1). Part of this work was presented at the 46th Annual Meeting of the Psychonomic Society, Toronto, Canada.

Appendix

Critical item pairs				Filler item pairs			
BACK	967	BAG	42	BUCK	20	BUG	4
HACK	3	HAG	–	CHUCK	14	CHUG	–
JACK	92	JAG	1	DOCK	8	DOG	75
KNACK	4	NAG	–	JOCK	–	JOG	–
LACK	110	LAG	3	LEAK	2	LEAGUE	69
RACK	9	RAG	10	LUCK	47	LUG	2
SACK	8	SAG	4	PLUCK	2	PLUG	23
SHACK	1	SHAG	1	PUCK	–	PUG	–
SNACK	6	SNAG	3	SMOCK	–	SMOG	1
STACK	9	STAG	8	TUCK	2	TUG	3
TACK	4	TAG	5	WICK	4	WIG	1

Note: When available, word frequency, as reported in Kučera and Francis (1967), is indicated.

References

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior Research Methods, 37*, 379–384.

Boersma, P., & Weenink, D. (2005). Praat: Doing phonetics by computer (Version 4.3) [Computer program]. Retrieved January 26, 2005. Available from http://www.praat.org/.

Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *The Journal of Acoustical Society of America, 116*, 3647–3658.

Dahan, D., & Gaskell, M. G. (2007). Temporal dynamics of ambiguity resolution: Evidence from spoken-word recognition. *Journal of Memory and Language, 57*, 483–501.

Dahan, D., Magnuson, J. S., & Tanenhaus, M. K. (2001). Time course of frequency effects in spoken-word recognition: Evidence from eye movements. *Cognitive Psychology, 42*, 317–367.

Eisner, F., & McQueen, J. M. (2005). The specificity of perceptual learning in speech processing. *Perception & Psychophysics, 67*, 224–238.

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*, 1166–1183.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review, 105*, 251–279.

Haijang, Q., Saunders, J. A., Stone, R. W., & Backus, B. T. (2006). Demonstration of cue recruitment: Change in visual appearance by means of Pavlovian conditioning. *Proceedings of the National Academy of Sciences of the United States of America, 103*, 483–488.

Hallé, P. A., & de Boysson-Bardies, B. (1994). Emergence of an early receptive lexicon: Infants' recognition of words. *Infant Behavior & Development, 17*, 119–129.

Hallet, P. E. (1986). Eye movements. In K. Boff, L. Kaufman, & J. Thomas (Eds.), *Handbook of perception and human performance* (pp. 10-1–10-112). New York: Wiley.

House, A. S. (1961). On vowel duration in English. *The Journal of Acoustical Society of America, 33*, 1174–1178.

Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In K. Johnson & J. W. Mullennix (Eds.), *Talker variability in speech processing* (pp. 145–166). San Diego: Academic Press.

Johnson, K. (2005). Speaker normalization in speech perception. In D. B. Pisoni & R. E. Remez (Eds.), *The handbook of speech perception* (pp. 363–389). Oxford, UK: Blackwell.

Kraljic, T., & Samuel, A. G. (2006). Generalization in perceptual learning for speech. *Psychonomic Bulletin & Review, 13*, 262–268.

Kučera, H., & Francis, W. H. (1967). *Computational analysis of present day American English*. Providence, RI: Brown University.

Labov, W., Ash, S., & Boberg, C. (2005). *The atlas of North American English*. New York: Mouton de Gruyter.

Ladefoged, P., & Broadbent, D. E. (1957). Information conveyed by vowels. *The Journal of Acoustical Society of America, 29*, 98–104.

Marslen-Wilson, W. D. (1973). Linguistic structure and speech shadowing at very short latencies. *Nature, 244*, 522–523.

Maye, J., Aslin, R. N., & Tanenhaus, M. K. (2008). The Weckud Wetch of the Wast: Lexical adaptation to a novel accent. *Cognitive Science, 32*, 543–562.

McQueen, J. M., Cutler, A., & Norris, D. (2006). Phonological abstraction in the mental lexicon. *Cognitive Science, 30*, 1113–1126.

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology, 47*, 204–238.

Nusbaum, H. C., Pisoni, D. B., & Davis, C. K. (1984). Sizing up the Hoosier Mental Lexicon: Measuring the familiarity of 20,000 words. *Research on Speech Perception Progress Report No. 10*, pp. 357–376.

Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics, 60*, 355–376.

Raphael, L. J. (1972). Preceding vowel duration as a cue to the perception of the voicing characteristic of word-final consonants in American English. *The Journal of Acoustical Society of America, 51*, 1296–1303.

Swingley, D. (2005). 11-month-olds' knowledge of how familiar words sound. *Developmental Science, 8*, 432–443.

Tanenhaus, M. K. (2007). Eye movements and spoken language processing. In R. Van Gompel, M. Fischer, W. Murray, & R. Hill (Eds.), *Eye movements: A window on mind and brain* (pp. 443–469). Oxford, England: Elsevier.

Zeller, C. (1997). The investigation of a sound change in progress: /ae/ to /e/ in Midwestern American English. *Journal of English Linguistics, 25*, 142–155.