8-2005

A Critical Evaluation of Visually Moderated Phonetic Context Effects

Lori L. Holt
Carnegie Mellon University, lholt@andrew.cmu.edu

Joseph D.W. Stephens
Carnegie Mellon University

Andrew J. Lotto
Boys Town National Research Hospital

Running Head:

VISUAL INFORMATION AND PHONETIC CONTEXT EFFECTS

A Critical Evaluation of Visually-Moderated Phonetic Context Effects

Lori L. Holt and Joseph D. W. Stephens

Carnegie Mellon University

Department of Psychology & Center for the Neural Basis of Cognition

Andrew J. Lotto

Boys Town National Research Hospital

Corresponding author:
Lori L. Holt
Department of Psychology
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213
Phone: 412/268-4964
Fax: 412/268-2798
Email: lholt@andrew.cmu.edu

Abstract

Fowler, Brown and Mann (2000) report a visually-moderated phonetic context effect in which a video disambiguates an acoustically ambiguous precursor syllable which, in turn, influences perception of a following syllable. The present experiments explore this finding and claims that stem from it. Experiment 1 failed to replicate Fowler et al. with novel materials modeled after the original study, but Experiment 2 successfully replicated the effect using Fowler et al.'s stimulus materials. This discrepancy was investigated in Experiments 3 and 4, which demonstrate that variation in visual information concurrent with the test syllable is sufficient to account for the original results. The Fowler et al. visually-moderated phonetic context effect appears to have been a demonstration of audiovisual interaction between concurrent stimuli and not an effect whereby preceding visual information elicits changes in the perception of subsequent speech sounds.

A Critical Evaluation of Visually-Moderated Phonetic Context Effects

Speech perception makes use of both auditory and visual information. McGurk and MacDonald (1976) demonstrated a potent and reliable effect in which mismatches between auditory and visual speech cues lead to perceptual identifications that would not be obtained based on the information presented to either modality alone. This effect has been replicated many times (see Colin & Radeau, 2003 for review) and is not disrupted by the perceiver's awareness of the audiovisual mismatch or on the perceiver's direction of attention to one modality or the other (Massaro, 1987).

Fowler, Brown, and Mann (2000) recently reported a visual influence on phonetic identification using an auditory/visual paradigm first described by Vroomen (1992, see also Green & Norrix, 2001). As implemented by Fowler et al., this paradigm takes advantage of a well-established demonstration of the context-dependent nature of speech perception (Mann, 1980) whereby listeners' identification of syllables ranging perceptually from /ga/ to /da/ is examined in the context of a preceding /al/ or /ar/ syllable. Perception is context dependent in that test syllables are more often identified as /ga/ when they are preceded by /al/. Listeners identify the same syllables more often as /da/ when they are preceded by /ar/. Fowler et al. investigated whether visual stimuli may moderate such effects of context on following syllables. Specifically, Fowler et al. created an acoustic stimulus that they judged to be perceptually ambiguous between /al/ and /ar/. This

stimulus served as the first-syllable soundtrack for a video of a male speaker articulating /alda/ or /arda/. Although the acoustically-ambiguous precursor remained constant across conditions and only the visual information varied to disambiguate the precursor, Fowler et al. (2000) observed a context effect on listeners' perception of subsequent acoustic syllables drawn from an acoustic /ga/ to /da/ series. Listeners identified test syllables more often as "ga" when visual information accompanying the precursor indicated an /al/ than when visual information indicated an /ar/.

The observation of a visually-moderated phonetic context effect indicates that auditory and visual information may be integrated to shift speech identification. At the most general level, this finding is not novel. In the classic McGurk effect, an audio soundtrack of /ga/ presented with a video stimulus of a face articulating /ba/ interact to create a perceived /da/. What is different about the Fowler et al. (2000) study is that visual information paired with the precursor modulates precursor perception which, in turn, may produce a shift in the identification of subsequent phonetic segments. This sort of context effect can be considered a "second order" or "indirect" effect of visual context. Similar second-order effects have been observed for other sources of information. Elman and McClelland (1988), for example, demonstrated that lexical information may serve to disambiguate an acoustically-ambiguous fricative between /s/ and /ʃ/ and, consequently, produce a context effect on perception of a member of a following

/t/-/k/ series (see also Pitt & McQueen, 1998; Magnuson, McMurray, Tanenhaus, & Aslin, 2003; Samuel & Pitt, 2003). The Fowler et al. paradigm is similar to that developed by Elman and McClelland in that it demonstrates that context effects can occur in the absence of acoustic change across precursor contexts. Apparently, either visual or lexical information can influence processing of an acoustically ambiguous syllable and thereby affect perception of a following syllable.

Since the Fowler et al. (2000) study, another investigation of visual moderation of phonetic context effects has been reported. In a study with methods very much like those of Fowler et al., Vroomen and de Gelder (2001) found little evidence of visually-moderated phonetic context effects. With unimodal acoustic stimuli, a preceding /s/ leads to more "ka" identifications for a /ta/ to /ka/ test syllable series than does a preceding /ʃ/ (Mann & Repp, 1981). Vroomen and de Gelder investigated whether an ambiguous acoustic fricative between /s/ and /ʃ/ would produce a phonetic context effect on a following /ta/ to /ka/ series when the fricative was disambiguated by video of a speaker saying /aska/ or /aʃka/. Although the video served to reliably influence participants' identification of the fricative, the resulting cross-modal percept did not shift identification of the following test syllable. This finding persisted even when reduced amplitude of the acoustic stimuli encouraged listeners to rely more on visual information.

Thus, although Vroomen and de Gelder (2001) and Fowler et al. (2000) employed very similar methods to examine the indirect, or second-order, auditory-visual context effects, their results were strikingly different. Whereas Fowler et al. observed a significant visually-moderated context effect on subsequent speech identification, Vroomen and de Gelder found no evidence of such effects. Given the close similarity of the two studies, the discrepancy in their findings is puzzling.

Indirect context effects arising from disambiguating lexical information have been very influential in understanding the dynamics of processing in spoken word recognition (e.g., Elman & McClelland, 1988; Pitt & McQueen, 1998; Magnuson et al., 2003; Samuel & Pitt, 2003). Indirect effects arising from visual information could similarly inform models of speech perception and spoken word recognition. Thus, we undertook to replicate the visually-moderated phonetic context effect reported by Fowler et al. in an attempt to determine the variables that are responsible for the presence or absence of these effects.

Experiment 1

In Experiment 1, we attempted to replicate the visually-moderated context effect reported by Fowler et al. (2000) using novel stimulus materials based on the Fowler et al. stimulus description.

Method

Participants

Twenty-two undergraduate students at Carnegie Mellon University participated. All were monolingual English speakers who reported normal hearing. The experiment took less than one hour and participants received course credit.

Materials

Figure 1 About Here

A caricature of the stimuli is shown in Figure 1. A series of ten acoustic test syllables ranging perceptually from /ga/ to /da/ was created using the cascade branch of the Klatt (1980) synthesizer according to the parameters specified by Lotto and Kluender (1998). These synthesis parameters are listed in Table I.

Table I About Here

Each member of the test-syllable series consisted of an 80-ms formant transition followed by a 170-ms steady-state segment. These stimuli comprised the set of stimuli indicated in Figure 1b. Each test syllable was paired with a vowel-consonant precursor with acoustic characteristics between /al/ and /ar/ (Figure 1a).

The precursor syllable consisted of a 200-ms steady-state segment followed by a 250-ms transition for a total length of 450 ms. Synthesis parameters (as specified by Fowler et al.) are shown in Table I. This relatively long stimulus duration was used by Fowler et al. to match the duration of the hyperarticulated syllables of the videotaped speaker. All syllables were synthesized with 16-bit resolution at a 10-kHz sampling rate. Test and precursor syllables were matched in RMS amplitude. The syllables were digitally appended with 50 ms of silence separating them and converted to PCM .WAV format for use as video soundtracks.

Two videos were created by digitally recording the face of a speaker (JDWS) as he spoke aloud the syllables /al/ and /ar/ in isolation. The speaker listened to the synthesized precursor sound while producing these syllables in order to produce utterances of approximately the same length as the precursor. The speaker produced several tokens of each syllable. One good visual token of each syllable was selected for use in the experiment (Figure 1c). These videos differed from the video materials of Fowler et al. (2000) in that video accompanied only the precursor (Figure 1c) and not the test syllable (Figure 1d). Although only the video information corresponding to /al/ and /ar/ is fundamental to the arguments advanced by Fowler et al., video corresponding to articulation of the final syllable was included in the original study. We discuss the implications of this stimulus factor in Experiments 3 and 4.

Each of the ten acoustic test syllables was combined with each of the two precursor videos to create a set of 20 audiovisual stimuli. The acoustic and optic portions of the stimuli were combined such that the synthesized acoustic precursor syllable was synchronized with each of the videos by matching its onset to the onset of the speaker's voice in the original video soundtrack. The original soundtrack was then removed, leaving only the synthesized acoustic stimulus. The video was cut at the offset of the precursor syllable, so that no visual information was present during the test syllable. During this period, a black box covered the region of the computer monitor that otherwise displayed the face articulating /al/ or /ar/.

Each video was trimmed from the beginning to equalize the silent interval before optic and acoustic onset of the syllables. The audiovisual stimuli were saved as Windows .AVI movies (30 frames/sec). Participants viewed the movies from a distance of approximately 31 inches with a monitor resolution of 800 x 600 pixels. The on-screen size of the movies was 640 x 480 pixels. Acoustic presentation was controlled by a Creative Audio PCI ES1370 sound card and participants heard acoustic stimuli diotically over Sennheiser HD-265 linear headphones at a comfortable listening level.

Procedure

During the experiment, participants sat in a sound-attenuated booth and listened over headphones while watching a LCD computer monitor mounted at

eye-level on the front wall of the booth. Before performing the task, participants were instructed to attend to both the sound and the video and to report what they heard. After viewing each audiovisual stimulus, participants indicated responses using an electronic button box labeled *AL-GA*, *AL-DA*, *AR-GA*, and *AR-DA*. Following each response, a prompt on the monitor directed participants to make an additional button press in order to proceed to the next trial.

Each experimental session began with a brief warm-up phase during which participants viewed four videos (/al/ and /ar/ audiovisual stimuli paired with acoustic endpoints drawn from the /ga/ to /da/ series). Each video was presented three times, in random order. Participants identified the syllables without feedback. This warm-up phase differed from that of Fowler et al. (2000), which began with a 100-item listening task in which listeners identified isolated test syllables as “ga” or “da” with no video accompanying the syllables.

After the warm-up task, participants performed an identical task with all 20 stimuli. Stimuli were presented in blocks of 20 and presentation order was randomized within each block. There were 20 blocks over all, for a total of 400 trials. As in the Fowler et al. (2000) study, participants did not receive feedback.

Results

To evaluate the influence of visual information on precursor identification, each participant’s accuracy in precursor syllable identification (as defined by the congruence of precursor identification to the actual identity of the syllable

articulated in the accompanying video) was computed. Figure 2 displays each participant's proportion of "al" responses when the acoustically-ambiguous precursor accompanied visual /al/ (congruent responses) versus visual /ar/ (incongruent responses). Open symbols that lie near the diagonal line represent participants whose responses to the precursor were relatively unaffected by visual information. Participants whose responses to the precursor were consistent with visual information are illustrated with filled symbols clustered in the upper-left corner of the figure. These individuals form a distinct subset of participants that is well-captured by a 65% correct criterion. Thus, participants whose precursor responses were less than 65% correct were excluded from further analyses because their performance showed little or no influence of visual information on precursor identification and such interaction is fundamental to the phenomenon.

Figure 2 About Here

In Experiment 1, only 9 out of 22 participants (40.9%) satisfied this criterion.[1] Among included participants, average accuracy in precursor identification was 82.8% for /al/ and 91.8% for /ar/. Listeners' accuracy in identifying precursors was worse than that reported by Fowler et al. (2000), who excluded no participants and yet reported 93.4% accuracy for /al/ and 84.0% accuracy for /ar/.

Among participants who exhibited a robust influence of visual information on precursor identification, there was little resulting effect on test-syllable identification. Figure 3 displays test-syllable identification results for above-criterion participants. Average percent “ga” response in each of the audiovisual conditions is plotted as a function of the test-syllable series. The visually-moderated phonetic context effect reported by Fowler et al. (2000) is clearly absent. In a 2 (audiovisual precursor condition) x 10 (test series syllable) ANOVA, only the main effect of test series member was significant, $F(9,72) = 14.7$, $p < .001$, indicating that participants labeled test syllables differently along the /ga/ to /da/ series. Contrasting with the findings of Fowler et al., the visual precursors had no significant influence on test-syllable identification, $F(1,8) = 3.20$, $p = .11$. For the data points where there appears to be a slight influence of visual precursor upon identification, the trend was opposite the predicted influence. There was no interaction between test series member and audiovisual condition ($F<1$).

Figure 3 About Here

Although the data entered into the analyses above were drawn only from participants exhibiting an influence of visual stimulus upon precursor identification, even these participants were not perfectly accurate in their

precursor perception. To more finely assess these data, an additional analysis included only trials for which precursor identification was congruent with the actual visual information. In this analysis, too, only the main effect of test series was observed, $F(9,72)=14.81$, $p<.001$. There was no significant effect of audiovisual precursor condition, $F(1,8)=3.57$, $p=.096$, and the very small trend was opposite the predicted influence of visual precursor. There was no interaction between test series member and audiovisual condition ($F<1$).

A final analysis explored whether participants' *responses* to the precursor syllables had any influence on identification of the test syllable. In this analysis, responses to test syllables were grouped according to how the precursor syllable was identified, regardless of the actual visual information presented on each trial. This analysis did not differ from the others. As in the previous analyses, only a main effect of test series member was observed, $F(9,72)=14.82$, $p < .001$. Again, there was no significant effect of audiovisual precursor condition, $F(1,8) = 2.87$, $p=.13$. There was no interaction of test series member and condition ($F<1$).[2]

Discussion

Experiment 1 failed to replicate the visually-moderated phonetic context effect reported by Fowler et al. (2000). One may question the robustness of the phenomenon given the failure of Vroomen and de Gelder (2001) to find evidence of a visually-moderated phonetic context effect and the null effect of Experiment 1. However, some differences existed between the materials and procedure used

in Experiment 1 and the description of methods in the study by Fowler et al. Most notably, stimuli of Experiment 1 included no visual information during presentation of the test syllable whereas, in the Fowler et al. study, visual information of a face articulating /da/ was present for each test syllable presentation across both conditions (a stimulus change reflecting differences in Figure 1d). Additionally, participants in Experiment 1 completed twice as many audiovisual trials as participants in the experiment of Fowler et al., but did not perform the initial audio-only listening task of Fowler et al.

Fowler et al. (2000) contend that the shifts in test-syllable identification are driven by the disambiguating precursor videos. It is difficult to reconcile how the minor differences between the present experiment and that of Fowler et al. might disrupt observation of the context effect. Experiment 2 is an attempt to replicate Fowler et al. with strict adherence to their procedures and stimulus materials.[3]

Experiment 2

Method

Participants

Fifteen students from Carnegie Mellon University served as participants. All were monolingual English speakers who reported normal hearing. The experiment took less than one hour and participants received course credit.

Materials

Audiovisual stimulus materials for Experiment 2 were identical to those of Fowler et al. (2000, Experiment 3b). An ambiguous acoustic precursor syllable and acoustic test syllables ranging from /ga/ to /da/ were paired with video of a speaker producing either /alda/ or /arda/. A set of isolated test syllables was created by digitally excising the test syllables from the audio track of the Fowler et al. stimuli. The acoustic stimuli were recorded with 16-bit resolution at a sampling rate of 22.2 kHz. The video stimuli were 320 x 240 pixels in size.

Procedure

Experiment 2 protocol matched the procedures used by Fowler et al. (2000). Participants first completed a 100-item listening task in which they identified isolated test syllables as either “ga” or “da” by pressing labeled buttons on an electronic response box. Acoustic presentation was under the control of Tucker Davis Technologies (TDT) System II hardware; stimuli were converted from digital to analog, low-pass filtered at 4.8 kHz, amplified and presented diotically over linear headphones (Sennheiser HD-265) at approximately 65-70 dB SPL(A). Stimuli were presented in ten blocks, and order of presentation was randomized within each block.

In the second part of the experiment, participants listened to disyllables while watching the speaker’s face articulate either /alda/ or /arda/ on a computer screen at a distance of approximately 31 inches. After each video, four response options appeared on the screen and participants used a button box labeled *AL-GA*,

AL-DA, *AR-GA*, and *AR-DA* to indicate the perceived disyllable. An additional button press was required to advance to the next trial. The audiovisual portion of the experiment consisted of 200 trials. Acoustic presentation was via a Creative Audio PCI ES1370 sound card. Stimuli were presented in blocks of 20 and order of presentation was randomized within each block.

Results

As in Experiment 1, data from participants with greater than 65% correct identification of the precursors were included in analyses (N=11, 73% of total). Average accuracy of precursor identification for these above-criterion participants was 76.8% for /al/ and 87.1% for /ar/. Accuracy for precursors was much better in Experiment 2 than Experiment 1, but still worse than reported by Fowler et al.

Figure 4 About Here

Figure 4 displays average test-syllable identification results for participants meeting criterion, illustrating that Experiment 2 successfully replicated the finding of Fowler et al. (2000). Listeners consistently labeled test syllables as “ga” more often when the acoustically-ambiguous precursor was accompanied by visual /al/ than when it was accompanied by visual /ar/. A 2 (audiovisual precursor condition) x 10 (test series member) ANOVA of participants’ percent “ga” responses in the audiovisual portion of the experiment

revealed significant main effects of condition, $F(1,10) = 7.82$, $p < .05$, and test series member, $F(9,90) = 36.27$, $p < .001$, and no significant interaction, $F<1$. The results were qualitatively the same when only trials for which listeners identified the precursor syllables congruent with video information were included.

In an additional analysis, responses to test syllables were grouped according to participants' responses to precursors. If precursor perception causes a context effect on perception of test syllables, the effect might be more reliably observed when analyzed in terms of participants' precursor identifications rather than the stimulus characteristics of the precursor. However, in a 2 (precursor response) x 10 (test series member) ANOVA, there was only a trend in the predicted direction, $F(1,10) = 3.90$, $p = .08$. Thus, the proportion of "ga" responses on trials in which participants had identified the precursor as /al/ was only marginally greater than on trials in which participants had identified the precursor as /ar/.

Discussion

Experiment 2 successfully replicated the major finding of Fowler et al. (2000) using stimulus materials from the original study. Participants identified test syllables as "ga" more often when the acoustically-ambiguous precursor was accompanied by visual /al/ than when the precursor was accompanied by visual /ar/, suggesting that phonetic context effects may occur even when acoustic information remains constant across conditions.

Why should preceding visual information have influenced perception in Experiment 2, but not in Experiment 1? Experiment protocols differed slightly across the studies, but acoustic and optic stimulus differences are more likely to have contributed to the disparity in findings across the two experiments. One difference in stimulus materials is that the speaker videotaped by Fowler et al. produced hyperarticulated versions of the precursor syllables whereas the speaker of Experiment 1 produced the /al/ and /ar/ with a more natural articulation.

The video accompaniment to the test syllables also differed across experiments. No visual information was present during the test syllables of Experiment 1 (a black screen replaced the face), whereas the original Fowler et al. (2000) test syllables were paired with video of the articulation of a syllable. For these latter stimuli, a speaker articulated /da/ in both conditions (i.e., /alda/ and /arda/). Ideally, the design of an experiment to test the second order influence of precursor video upon perception of test syllables should eliminate or control the test-syllable video across conditions for which precursor video varies. In the Fowler et al. video, small movements in the speaker's head would cause unnatural jumps in the image at the syllable break should the same video clip of /da/ be appended to both /al/ and /ar/ precursor videos. As a result, Fowler et al. deemed it infeasible to use identical second-syllable videos across conditions. Instead, Fowler et al. videotaped a male articulating /alda/ and /arda/.

To examine the role of the test-syllable videos in test-syllable identification, Fowler et al. (2000) conducted a short video-only pilot study in which participants labeled the second syllable in the videotapes. With no audio track, participants' "da" responses were approximately equal across /alda/ and /arda/ video contexts (67% versus 65%, respectively in a forced-choice task). Fowler et al. took this as evidence that although the video accompanying the test syllable varied across conditions, it was *perceptually* equivalent. Thus, Fowler et al. argued, the test-syllable video should not affect test syllable perception differentially across conditions. Nonetheless, frame-by-frame inspection of the video clips reveals that the visual information present during the test syllables differed quite significantly across conditions. Figure 5 illustrates the first frame of the video accompanying acoustic onset of test syllables in the Fowler et al. stimulus materials. The left panel displays the first moment of articulation of the /da/ in /alda/. The panel on the right displays the /da/ as produced in /arda/. Clearly, the two are not identical.

Figure 5 About Here

It is possible that the video-only identification experiment reported by Fowler et al. (2000) may not have been an adequate index of the perceptual influence of video accompanying test syllables because it did not consider the

potential for auditory and visual cues to interact in perception. In the classic McGurk effect, visual and auditory cues cooperate to create a percept that is different from that indicated by either of the cues in isolation (McGurk & MacDonald, 1976). Experiments 3 and 4 investigate the ramifications of the subtle differences in stimulus materials present across Experiments 1 and 2 with two separate stimulus manipulations to the original materials of Fowler et al. Experiment 3 investigates the effect of precursor video upon identification of the test syllable by replacing the test-syllable video with a black box (as in Experiment 1). Experiment 4 investigates the impact of subtle differences across the two test-syllable videos (as shown in Figure 5) upon listeners' phonetic identification by removing the precursor audio and video in its entirety and testing for an effect of test-syllable video upon test-syllable identification.

Experiment 3

Experiment 3 uses the stimulus materials of Fowler et al. (2000) with a black box occluding the video concurrent with the second syllable as in Experiment 1. If a second order influence of precursor video is responsible for the results of Experiment 2, then Experiment 3 should replicate the observation of a visually-moderated phonetic context effect. However, if the differences in video present during the test syllable are responsible for the effect observed in Experiment 2 then Experiment 3 should produce results similar to those of

Experiment 1 for which no visually-moderated phonetic context effect was observed.

Method

Participants

Participants were 17 individuals recruited from the Pittsburgh, PA area through campus fliers and electronic bulletin board advertisements. All were monolingual English speakers who reported normal hearing. The experiment took less than one hour and participants were paid $7.

Materials

Audiovisual stimuli for this experiment were modified versions of the stimuli used by Fowler et al. (2000). Using Adobe Premiere (Adobe Systems, San Jose, CA), the videos were cut 200 ms before the onset of the test syllables in the corresponding soundtracks.[4] All video frames after that point in time were discarded, leaving only a blank screen during presentation of the test syllables.

Procedure

The experimental protocol was identical to that of Experiment 2.

Results

As in Experiments 1 and 2, participants with greater than 65% correct identification of the precursors were included in analyses (N=12, 71% of total). Average accuracy of precursor identification for these above-criterion participants

was 77.7% for /al/ and 80.8% for /ar/. Mean identification responses for these participants are illustrated in Figure 6.

Figure 6 About Here

A 2 (audiovisual precursor condition) x 10 (test series member) ANOVA of percent "ga" responses in the audiovisual portion of the experiment revealed a significant main effect of test series member, $F(9,99) = 79.67$, $p < .001$, no significant effect of condition, $F(1,11) = 1.05$, $p = .33$, and no significant interaction, $F(9,99) = 1.44$, $p = .18$. There was also no effect of audiovisual precursor when only the trials in which participants correctly identified the precursor were analyzed, $F(1,11) < 1$. In an additional analysis, responses to test syllables were grouped according to participants' responses to precursors. In a 2 (precursor response) x 10 (test series member) ANOVA, there was also no effect of precursor response, $F(1,11) < 1$.

Experiment 2 provided evidence of a robust phonetic context effect using the stimulus materials of Fowler et al. (2000). These results were presumed to be a visually-moderated phonetic context effect arising from the phonetically-disambiguating precursor video. The results of Experiment 3 suggest a different interpretation. When visual information coincident with test-syllables is eliminated by replacing the test-syllable video with a black box, no effect of

precursor visual context is observed. It is possible that small differences in video accompanying the test-syllable across conditions may produce the effect of context replicated in Experiment 2.

Experiment 4

In Experiment 4, the original stimulus materials of Fowler et al. (2000) were presented without optic and acoustic precursors (Figure 1a and 1c). This manipulation provided a strict test of the perceptual influence of visual information paired with test syllables. It also addresses the concern that the original video-only pilot test of Fowler et al. did not adequately assess perceptual equivalence of the video accompanying the second syllables in the /alda/ and /arda/ videos. Participants listened to test syllables while viewing either the visual /da/ excised from /alda/ or visual /da/ taken from /arda/. If test-syllable identification is influenced by the video condition then it is reasonable to propose that the phonetic labeling shift observed in Experiment 2 arose from contemporaneous rather than preceding visual information.

Method

Participants

Participants were 14 volunteers recruited from the Pittsburgh, PA area through campus fliers and electronic bulletin board advertisements. The experiment lasted approximately one-half hour and participants were paid $5. All participants were monolingual English speakers who reported normal hearing.

Materials

Audiovisual stimuli for Experiment 4 were modified versions of the stimuli used in Fowler et al. (2000) and Experiments 2 and 3. Using Adobe Premiere software, the original Fowler et al. video clips were cut at the beginning of the frame in which acoustic information for the test syllable began such that the frames illustrated in Figure 5 were the first frames of the video stimulus. Video that had preceded the onset of the test syllable was discarded (Figure 1c). The acoustic stimuli from Experiment 2 were cut synchronously with the beginning of the video frame in which the acoustic information for the test syllable began and acoustic information for the precursor was eliminated (Figure 1a). The resulting audiovisual stimuli therefore consisted of acoustic test syllables paired with either optic /da/ from /alda/ or optic /da/ from /arda/ (Figure 1b and 1d).

Procedure

The procedure for Experiment 4 was identical to the procedure for Experiment 2, except that “ga” and “da” were the only response options.

Results

Results of Experiment 4 are presented in Figure 7. There was a significant influence of the concurrent visual stimulus upon test-syllable identification. Specifically, listeners responded “ga” more often when test syllables accompanied visual /da/ from the /alda/ video than when they accompanied visual /da/ from the /arda/ video. Qualitatively, this influence is the same as that found in Experiment

2 and reported by Fowler et al. despite the fact that the Experiment 4 stimuli lacked both acoustic and optic precursors. A 2 (audiovisual test-syllable condition) x 10 (test series member) ANOVA revealed significant main effects of audiovisual test-syllable condition, $F(1,11) = 12.4$, $p < .005$, and test series member, $F(9,99) = 106.8$, $p < .001$, and a significant interaction, $F(9,99) = 10.4$, $p < .001$. An ANOVA comparing effect size across Experiment 2 and Experiment 4 found no significant difference, $F<1$.[5] Given that the acoustic stimuli were constant across the conditions and that there were no auditory or visual precursors, differences across the two visual /da/ stimuli appear to have been the basis of the observed effect.

Figure 7 About Here

Discussion

In Experiment 4, precursor syllables were eliminated entirely, leaving neither visual nor auditory context stimuli. Even without a precursor stimulus, participants' responses to the test syllables shifted as a function of the accompanying /da/ videos. Although Fowler et al. (2000) found no significant difference in participants' labeling of the second syllable of the /alda/ and /arda/ videos in a video-only condition, there appears to be a rather significant influence of the two visual /da/ tokens when they are allowed to interact with the acoustic

test syllables. In all, the significant influence of test-syllable video upon test-syllable perception greatly tempers the claims that can be made about the existence of indirect, or second-order, visually-moderated phonetic context effects. The effect observed in Experiment 2 and by Fowler et al. appears to arise from integration of concurrent auditory and visual information.

General Discussion

The experiments presented here were undertaken to examine the discrepancy in the literature regarding visually-moderated phonetic context effects whereby the visual information concurrent with an acoustically ambiguous precursor modulates perception of a following speech signal. Whereas Fowler et al. (2000) report that visual precursor information can produce a shift in listeners' identification of a following test syllable, Vroomen and de Gelder (2001) observe no such effect despite using a very similar paradigm. Experiment 1 was undertaken to attempt to replicate the observations of Fowler et al. However, no effect of visually-moderated phonetic context was observed with novel stimulus materials. In Experiment 2, a full replication of the Fowler et al. methods using the original Fowler et al. stimulus materials resulted in observation of a significant effect of video on test syllable identification.

Two additional experiments were conducted to examine the discrepancy in effects of the stimulus materials; these experiments manipulated the original Fowler et al. stimuli in a complementary manner. In Experiment 3, the video

concurrent with the test syllable presentation was deleted and replaced with a black screen such that visual information was present only during the acoustically-ambiguous precursor. This manipulation of the Fowler et al. materials resulted in a stimulus design matched to that of Experiment 1. As was observed in Experiment 1, there was no effect of precursor visual information on test syllable identification. The failure to observe visually-moderated context effects in Experiments 1 and 3 led to the hypothesis that the visual effects reported by Fowler et al. were due to differences in the video coinciding with test syllable presentation. This hypothesis was investigated in Experiment 4 by deleting the audio and video precursors from the Fowler et al., leaving only the test syllable video and audio intact. Listeners exhibited a shift in identification as a function of whether the /da/ video originated from the Fowler et al. /alda/ or /arda/ materials. This shift was statistically equivalent to that obtained in Experiment 2 with intact precursor audio and video. Thus, the effect of visual information reported by Fowler et al. appears to arise from the influence of *concurrent* rather than *preceding* video. The present results are compatible with the findings of Vroomen and de Gelder (2001), who observed that although listeners integrated concurrent auditory and visual information, integration did not have an influence on perception of a following syllable. The discrepancy between the observations of Fowler et al. and Vroomen and de Gelder appear to be due to the fact that the video accompanying the test-syllable audio in the materials of

Fowler et al. were not matched across conditions. Overall, the results of Fowler et al., Vroomen and de Gelder, and the present experiments are best accounted for by the influence of *concurrent* visual information on speech identification; there is no consistent evidence of a visually-moderated context effect when visual information disambiguates a preceding context syllable.

Theoretical Consequences

The classic demonstrations of non-auditory or cross-modal influences in speech perception, such as McGurk effects, may be referred to as *first order* interactions (McGurk & MacDonald, 1976). That is, there is an interaction of information specifying the same phonetic segment or syllable. For example, patterns of acoustic energy consistent with /da/ may interact with (typically concurrent) optic patterns of articulation that influence the perceived syllable. Such first order interactions may be accounted for by models of cross-modal integration that are general in nature (e.g., Massaro, 1987; 1998) or by models specific to speech perception (e.g., Liberman & Mattingly, 1985; see Miller, 1990). The lexical or statistical context effects described by Elman & McClelland (1988) and Pitt & McQueen (1998; see also Magnuson et al., 2003; Samuel & Pitt, 2003) are examples of *second order* interactions, with lexical or statistical information influencing the perceived identity of acoustically ambiguous context sounds which, in turn, influences the perception of a subsequent test syllable. Elman & McClelland's TRACE model (1988; McClelland and Elman, 1986)

accommodates these interactions by activation that flows from lexical representations to phonemic representations that can influence the activation of acoustic/phonetic features in subsequent time-slices.

The visually-moderated effects described by Fowler et al. (2000) appeared to be examples of second order interactions. The visual information disambiguated the perceived identity of the acoustically ambiguous context sounds and test syllable identification appeared to be moderated by the perceived context identity. However, the results of the four experiments presented here provide little evidence for second order interaction of visual and auditory information. In Experiments 1 and 3, the visual information concurrent with the context syllable was sufficient to influence its perceived identity. However, this had no effect on test syllable identification when no video accompanied the test syllable. In Experiment 2, when responses to the original stimuli of Fowler et al. were analyzed in terms of participants' identification responses to the context syllable (as opposed to the intended identity of the visual display) the context effect was non-significant. Finally, Vroomen and de Gelder (2001) failed to find evidence of second order interactions for very similar stimuli despite observing robust first-order interactions of concurrent auditory and visual information.

The lack of a strong second order interaction in the present experiments is unexpected given demonstrations of lexical second-order effects (Elman & McClelland, 1987; Magnuson et al., 2003; Samuel & Pitt, 2003). On the surface,

the two experimental paradigms share many similarities. In both cases the phonemic identity of an acoustic context syllable is disambiguated by non-acoustic information. When the disambiguating information is lexical or statistical there is a concomitant change in test syllable identity, but when the information is visual there is no attendant change in perceived test syllable identity. For models like TRACE that rely on the phonemic content of the context to predict effects on test syllable identification, this lexical-visual discrepancy is difficult to reconcile (McClelland & Elman, 1986). In TRACE, the activation of the phoneme representations is not differentiated by source; thus, there is no mechanism to distinguish context effects on subsequent test stimuli based on the source of disambiguating information.

Whereas there is a lack empirical evidence for second order interactions in auditory-visual speech recognition, the present experiments join many previous studies in demonstrating the robust interaction of concurrent visual and auditory information on speech perception (see Colin & Radeau, 2003 for review). In Experiment 4, although no context video and audio was presented, participants' responses were influenced by whether the video was from productions of /ar da/ or /al da/. The shift in responses was in the same direction as predicted by the /al/ and /ar/ contexts of the original full video. One possible explanation for this shift is that there was enough visual information present about the context syllable in the test syllable video to induce a context effect. Specifically, the differences in

the two videos at the onset of the test syllables were probably the visual consequence of carryover coarticulation from the context syllables. Thus, the video time-aligned to the test syllables may have carried visual coarticulatory information that specified the context syllable. This information may have been sufficient to shift test syllable identification, producing a second-order visual context effect. Within a gesturalist approach to speech perception (Fowler, 1986), the acoustic consequences of coarticulation would be parsed and attributed to the gestures of the context that caused the differences in the video. In this light, the concurrent auditory-visual context effect demonstrated in Experiment 4 could still be an example of a second-order visually-moderated context effect as suggested by Fowler et al. (2000).

There are several potential problems for this explanation of the auditory-visual context effect observed in Experiment 4. First, the acoustic and optic silence preceding the test syllables in this experiment was not just a lack of information about the context syllable; it was *information that the stimulus lacked a context syllable*. One would expect, at the least, that the effect of the visual information for the context syllable would be greater in conditions where it was fully present in auditory and visual domains (Experiment 2) than when only a small visual vestige (the visual coarticulatory influence) of the context syllable was present (Experiment 4). However, there was no difference in effect size between these two experiments. It is also difficult to reconcile why there was no

effect of context when the context video was present, but test syllable video was absent (Experiments 1 and 3). Presumably, the system would also attribute coarticulatory influences to the context syllables in these conditions where there was strong evidence for the presence of a context. The null effects in these experiments suggest that the effect of the visual stimulus in Experiment 4 was not because it specified the context syllable (a second order interaction) but because the differences in the video specified the identity of the test syllable (a first order interaction). That is, the concurrent visual information modulated participants' identification of the test syllable without influence of a preceding context.

Whether this first-order interaction is the result of speech-gesture specific perception (e.g., Fowler et al., 2000) or general cross-modal perceptual processes (e.g., Massaro, 1987; 1998) remains an open question. The mere presence of audio-visual interactions does not distinguish among theories of speech perception. All of the major theories acknowledge that information for recognition of speech sounds can come from multiple sources (e.g., lexical, visual). What the current results demonstrate is that these sources may differ in the extent of their interactions with acoustic information. Lexical and statistical/distributional information have been shown to result in second-order interactions, whereas there is thus far no evidence that visual information produces such indirect interactions. This distinction may provide an opportunity

to further test and clarify models of speech sound identification and word recognition.

References

Colin, C., & Radeau, M. (2003). The McGurk illusions in speech: 25 years of research. *Anee Psychologique, 103,* 497-542.

Elman, J. L., & McClelland, J. L. (1988). Cognitive penetration of the mechanisms of perception: Compensation for coarticulation of lexically restored phonemes. *Journal of Memory and Language, 27,* 143-165.

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

Fowler, C. A., Brown, J. M., & Mann, V. A. (2000). Contrast effects do not underlie effects of preceding liquids on stop-consonant identification by humans. *Journal of Experimental Psychology: Human Perception and Performance, 26,* 877-888.

Green, K. P., & Norrix, L. W. (2001). Perception of /r/ and /l/ in a stop cluster: Evidence of cross-modal context effects. *Journal of Experimental Psychology: Human Perception and Performance, 27,* 166-177.

Klatt, D. H. (1980). Software for a cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America, 67,* 971-990.

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

Lotto, A. J., & Kluender, K. R. (1998). General contrast effects in speech perception: Effect of preceding liquid on stop consonant identification. *Perception & Psychophysics, 60,* 602-619.

Magnuson, J. S., McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2003). Lexical effects on compensation for coarticulation: The ghost of Christmash past. *Cognitive Science, 27,* 285-298.

Mann, V. A. (1980). Influence of preceding liquid on stop-consonant identification. *Perception & Psychophysics, 28,* 407-412.

Mann, V. A., & Repp, B. H. (1981). Influence of preceding fricative on stop consonant perception. *Journal of the Acoustical Society of America, 69,* 548-558.

Massaro, D. W. (1987). *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry.* Hillsdale, NJ: Erlbaum.

Massaro, D. W. (1998). *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle*. Cambridge, MA: MIT Press.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18,* 1-86.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264,* 746-748.

Miller, J. L. (1990). Speech perception. In D. N. Osherson & H. Lasnik (Eds.), *An Invitation to Cognitive Science* (Vol. 1, pp. 69-93). Cambridge, MA: MIT Press.

Munhall, K.G., Gribble, P., Sacco, L., & Ward, M. (1996). Temporal constraints on the McGurk effect. *Perception & Psychophysics, 58,* 351-362.

Pitt, M. A., & McQueen, J. M. (1998). Is compensation for coarticulation mediated by the lexicon? *Journal of Memory and Language, 39,* 347-370.

Samuel, A. G., & Pitt, M. A. (2003). Lexical activation (and other factors) can mediate compensation for coarticulation. *Journal of Memory and Language, 48,* 416-434.

Vroomen, J. (1992). *Hearing voices and seeing lips: Investigations in the psychology of lipreading.* Unpublished doctoral dissertation. Tilburg University, Tilburg, The Netherlands.

Vroomen, J., & de Gelder, B. (2001). Lipreading and the compensation for coarticulation mechanism. *Language & Cognitive Processes, 16,* 661-672.

Author Note

Lori L. Holt and Joseph D. W. Stephens, Department of Psychology, Carnegie Mellon University and Center for the Neural Basis of Cognition, Pittsburgh, PA; Andrew J. Lotto, Boys Town National Research Hospital, Omaha, NE.

Empirical work in this paper was presented at the 43rd annual meeting of the Psychonomic Society. The research efforts of the second author were supported by the Center for the Neural Basis of Cognition and a National Defense Science and Engineering Graduate Fellowship. National Institutes of Health award R01 DC04674-01 to LLH and AJL also supported the project. The authors wish to thank Jay McClelland for thoughtful discussions of the work and Christi Adams, Siobhan Cooney, Ashley Episcopo and Seth Liber for their assistance in data collection.

Correspondence concerning this article should be addressed to Lori L. Holt, Department of Psychology, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213. Email: lholt@andrew.cmu.edu.

Endnotes

[1] This low inclusion rate may be a consequence of the magnitude of the visual difference in /al/ and /ar/ videos of Experiment 1. Whereas the speaker videotaped by Fowler et al. (2000) hyperarticulated the precursor syllables, the present speaker produced tokens with a natural articulation style. As a result, the two video precursors for this experiment may have been less visually distinct than those of Fowler et al. We explore these differences in discussion of Experiments 3 and 4. Participants excluded from the present analyses due to poor precursor accuracy responded /ar/ correctly 78.97% of the time, but responded /al/ correctly only 28.5% of the time. Thus, this sub-group of participants tended to hear precursors as /ar/ irrespective of video.

[2] Duplicate analyses were conducted including all 22 participants' responses (without regard to the inclusion criterion). The results are not qualitatively different from those observed for the participants with precursor responses that were more congruent with the actual video. There was a trend toward an effect of visual condition upon participants' test-syllable labeling responses, $F(1,21) = 3.78$, $p = 0.07$, but this trend was in the opposite direction of that observed by Fowler et al. (2000). There was no interaction ($F<1$). In an analysis of only trials for which participants correctly identified the precursor, there was a significant influence of visual stimulus upon test syllable perception,

but again in the direction opposite that observed by Fowler et al., $F(1,21) = 7.76$, $p = 0.01$. It is important to note, however, that this analysis may be misleading because participants with generally incongruent precursor responses (those excluded from the main analyses) had very few correct precursor trials, especially in the /al/ condition.

[3] We thank Carol Fowler and Julie Brown, who kindly provided the original stimuli used by Fowler et al. (2000).

[4] Munhall, Gribble, Sacco & Ward (1996) report McGurk effects when audio stimuli lags video by as much as 180 ms. To mitigate the concern that video immediately preceding the onset of the test syllable might remain effectively "concurrent" perceptually, the videos for the present experiment were cut 200 ms before the onset of the test syllables. Results confirm that this operation did not greatly impact participants' ability to identify the precursor; participants with above-criterion performance heard the precursors correctly 77.7% of the time for /al/ and 80.8% of the time for /ar/.

[5] Overall, the effect sizes of Experiment 2, M=8.27%, SE=2.96%, and Experiment 4, M=6.83%, SE=1.94%, were somewhat smaller than that reported by Fowler et al. (2000), M=11.1%.

Table I

Synthesis Parameters for Syllables Used in Experiment 1

	Vowel	Liquid	Stop	Vowel
Frequency	[a]	Ambiguous [l/r]	[g-d]	[a]
f0	110	110	110	110[a]
F1	750	556	300	750
F2	1200	1300	1650	1200
F3	2450	2150	1800-2700	2450
F4	2850	2850	2850	2850

Note. Parameters for the vowels indicate steady-state values. Parameters for the liquid indicate offset values after 250-ms linear transitions. Parameters for the stops indicate onset values before 80-ms linear transitions.

[a] *f*0 decreased linearly from 110 Hz to 95 Hz over the final 100 ms of each test syllable.

Figure Captions

Figure 1. A schematic representation of the stimuli used by Fowler, Brown, and Mann (2000) and investigated in the current studies. Captions labeled *a*, *b*, *c*, and *d* refer to the auditory precursor, auditory test syllable, visual precursor and visual test syllable, respectively. The bottom half of the figure schematizes the specific stimulus characteristics of each experiment.

Figure 2. Scatter plot of each listener's percent "al" responses when the visual stimulus corresponded to articulation of /al/ (Congruent Responses) versus the same listener's percent "al" responses when the visual stimulus corresponded to articulation of /ar/ (Incongruent Responses) in Experiment 1. Filled circles illustrate listeners with greater than 65% accuracy in labeling the precursor congruently with the articulation presented in the video. These listeners exhibit a strong influence of visual information on syllable identification. The remaining listeners' responses (open circles) were relatively unaffected by visual information.

Figure 3. Percent "ga" responses averaged across Experiment 1 participants meeting criterion for precursor identification performance (see text for inclusion criteria). Data are presented as a function of test syllable, varying across an acoustic dimension ranging perceptually from /ga/ to /da/. Filled circles indicate participants' identification of these syllables in the context of audiovisual /ar/ and filled triangles correspond to responses in the context of audiovisual /al/.

Figure 4. Percent "ga" responses averaged across Experiment 2 participants meeting criterion (see text for inclusion criteria). Data are presented as a function of test syllable, varying across an acoustic dimension ranging perceptually from /ga/ to /da/.

Filled circles indicate participants' identification of these syllables in the context of audiovisual /ar/ and filled triangles correspond to responses in the context of audiovisual /al/.

Figure 5. The first frames of the video accompanying acoustic onset of test syllables in the original Fowler et al. (2000) video stimulus materials. The photograph on the left is from the test syllable /da/ from /alda/. The photograph on the right is the first frame of the /da/ from /arda/.

Figure 6. Percent "ga" responses to audiovisual test syllables excised from the original Fowler et al. (2000) stimulus materials averaged across Experiment 3 participants. Data are presented as a function of test syllable, varying across an acoustic dimension ranging perceptually from /ga/ to /da/. Filled circles indicate participants' identification of these syllables in the context of the audiovisual /ar/ and filled triangles correspond to responses in the context audiovisual /al/. No video accompanied the test syllables.

Figure 7. Percent "ga" responses averaged across Experiment 4 participants. Data are presented as a function of test syllable, varying across an acoustic dimension ranging perceptually from /ga/ to /da/. There were no audio or video precursors in this experiment. Filled circles indicate participants' identification of test syllables when test-syllable video spliced from the Experiment 2 /ar da/ condition was presented concurrently with the audio test syllable. Filled triangles correspond to responses when the same audio test syllables were identified with concurrent video spliced from the test-syllable segment of the /al da/ condition of Experiment 2.

Figure 1

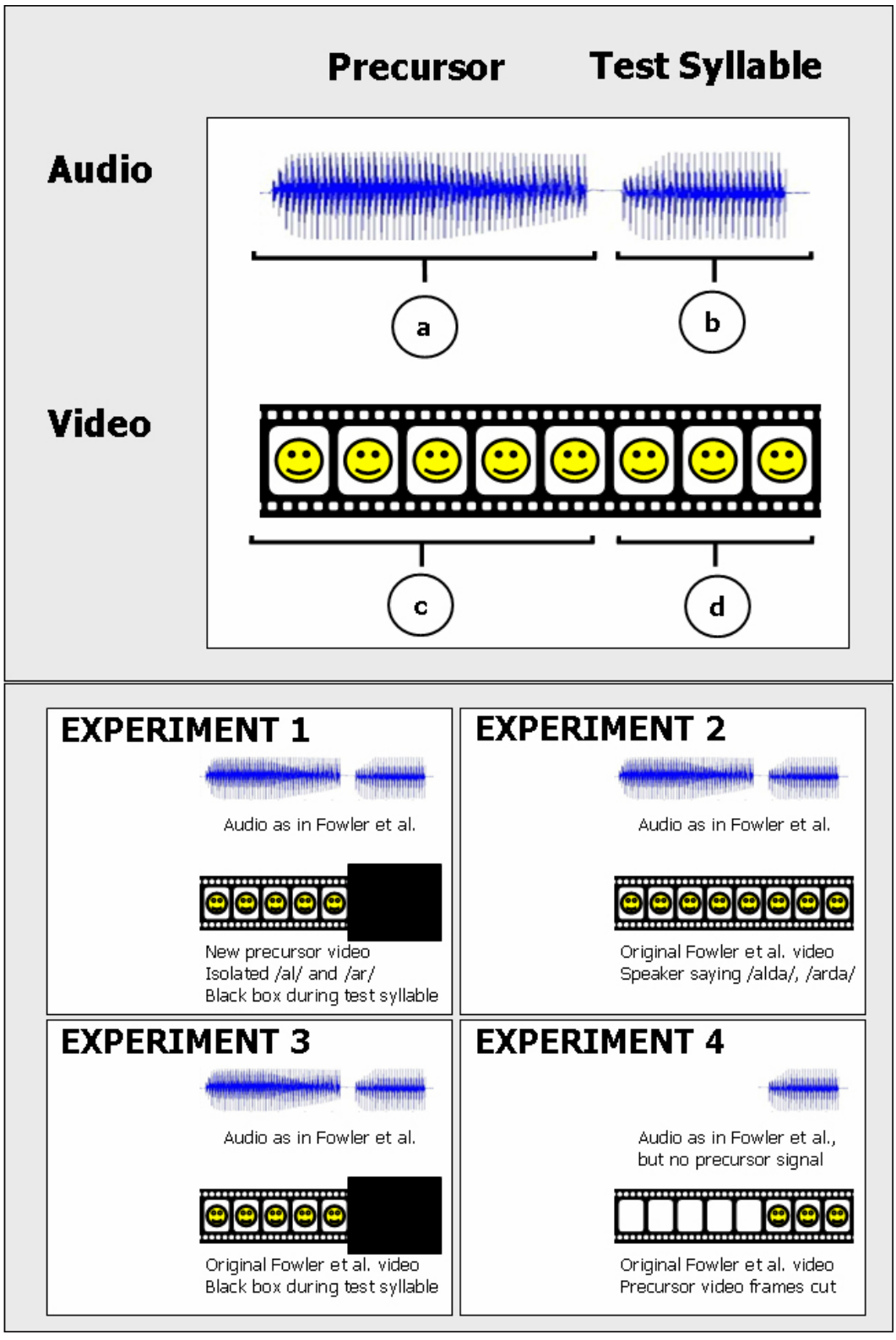

Precursor
Test Syllable
Audio
Video
a
b
c
d
EXPERIMENT 1
Audio as in Fowler et al.
New precursor video
Isolated /al/ and /ar/
Black box during test syllable
EXPERIMENT 2
Audio as in Fowler et al.
Original Fowler et al. video
Speaker saying /alda/, /arda/
EXPERIMENT 3
Audio as in Fowler et al.
Original Fowler et al. video
Black box during test syllable
EXPERIMENT 4
Audio as in Fowler et al.,
but no precursor signal
Original Fowler et al. video
Precursor video frames cut

Figure 2

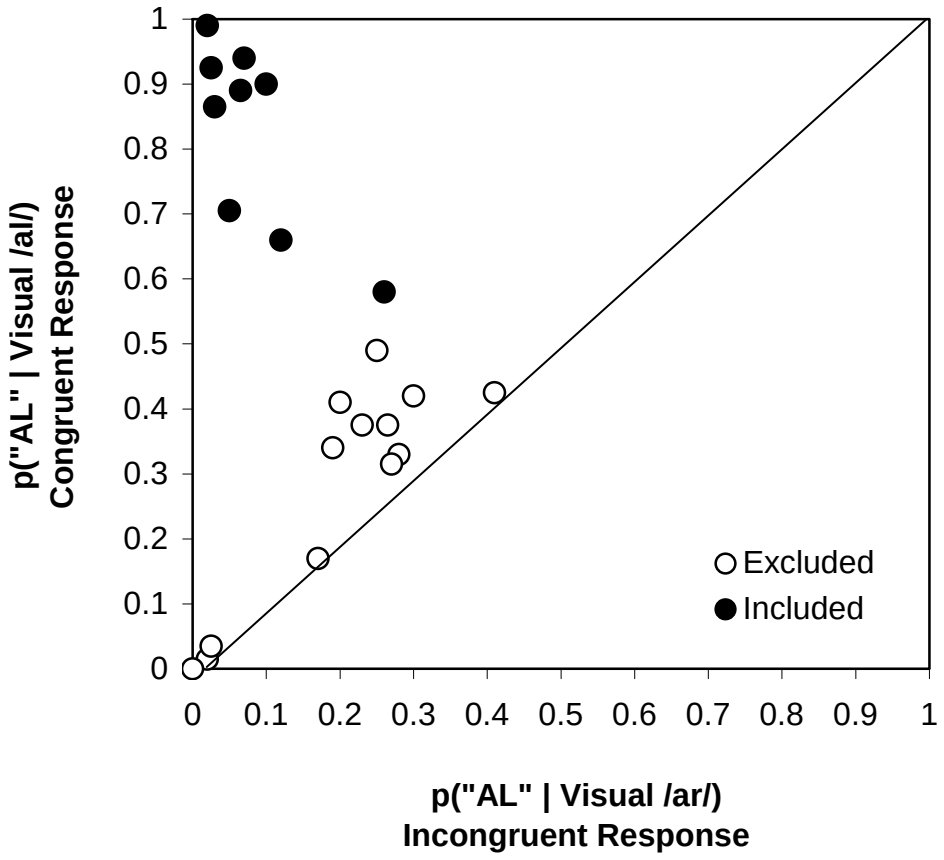

p("AL" | Visual /al/)
Congruent Response
p("AL" | Visual /ar/)
Incongruent Response
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Excluded
Included

Figure 3

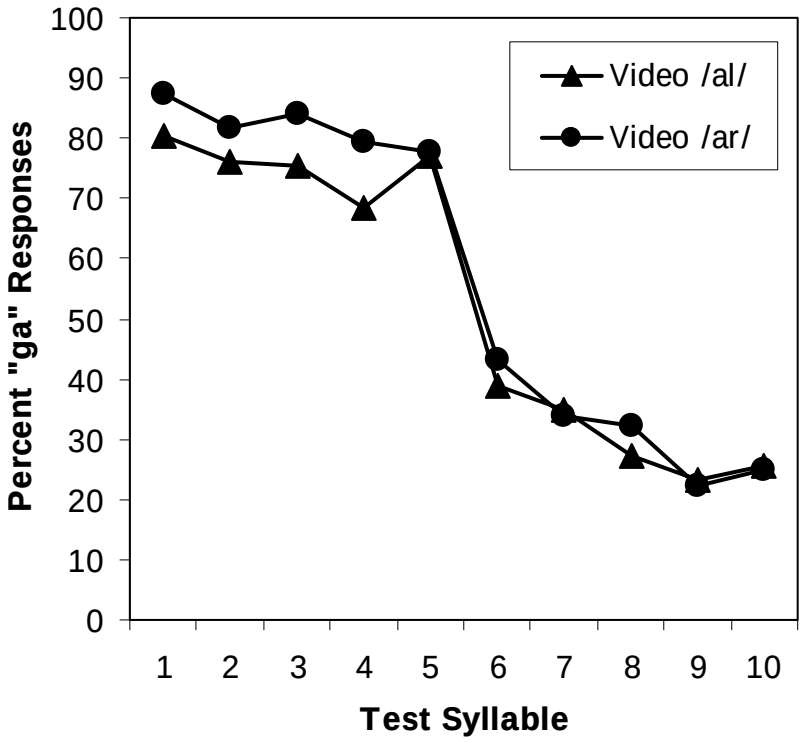

100
90
80
70
60
50
40
30
20
10
0
Percent "ga" Responses
1
2
3
4
5
6
7
8
9
10
Test Syllable
Video /al/
Video /ar/

Figure 4

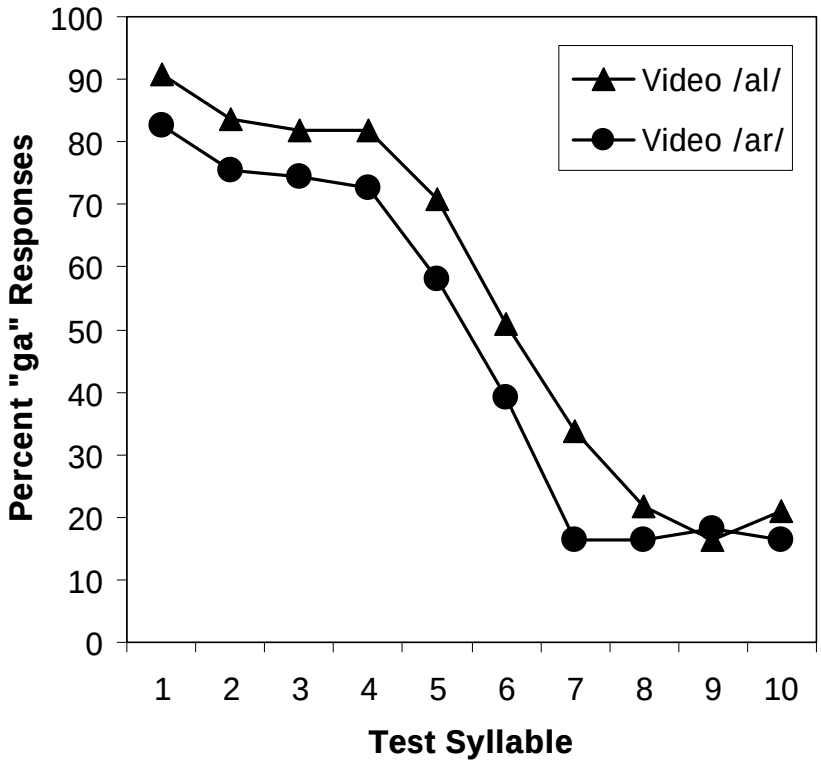

100
90
80
70
60
50
40
30
20
10
0
Percent "ga" Responses
1
2
3
4
5
6
7
8
9
10
Test Syllable
Video /al/
Video /ar/

Figure 5

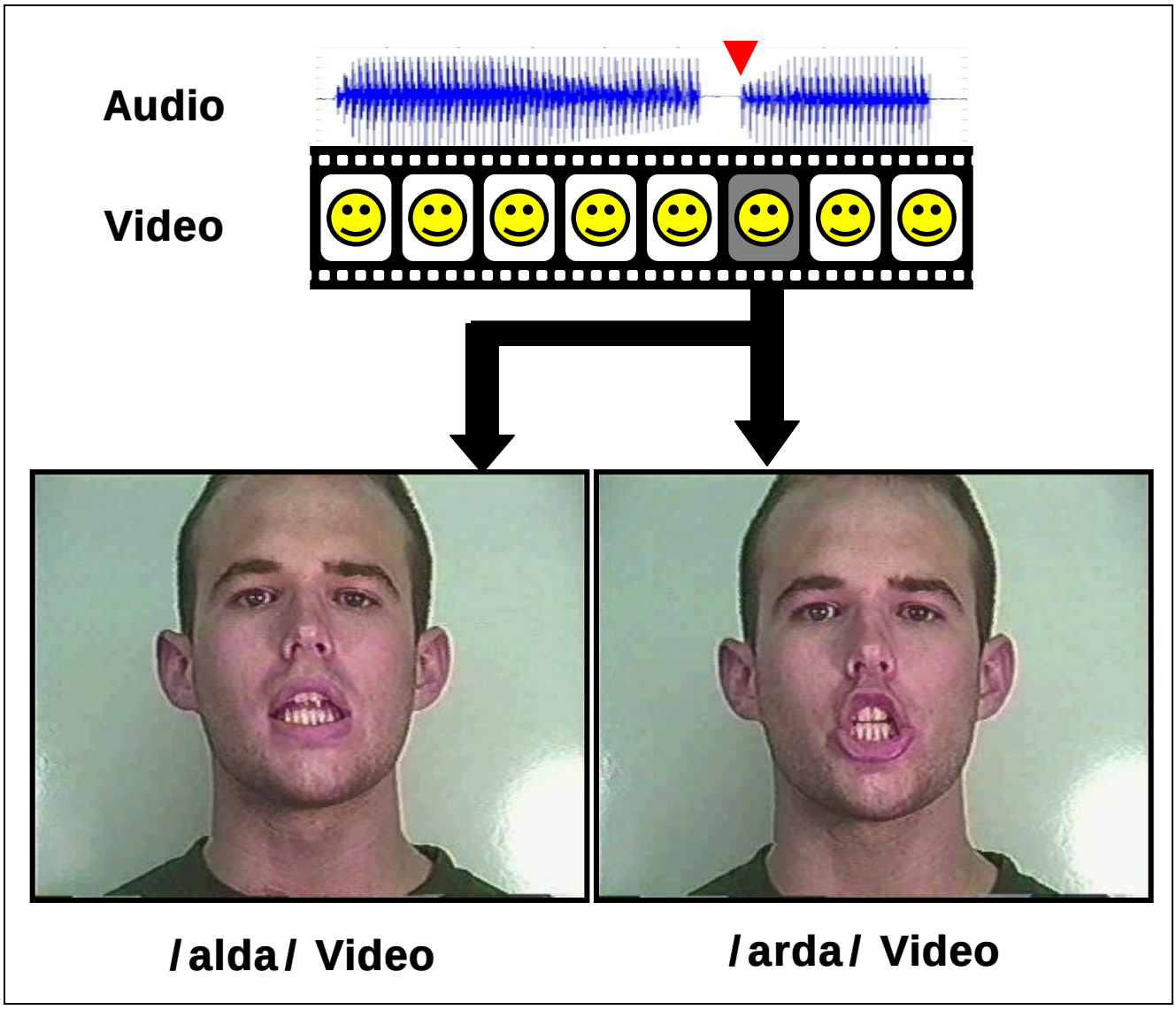


Figure 6

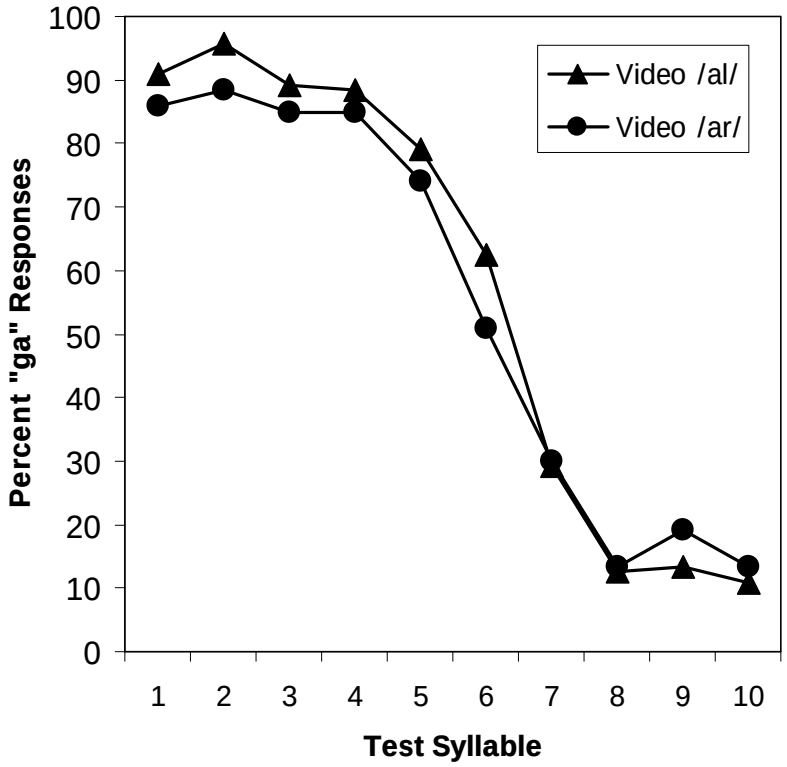
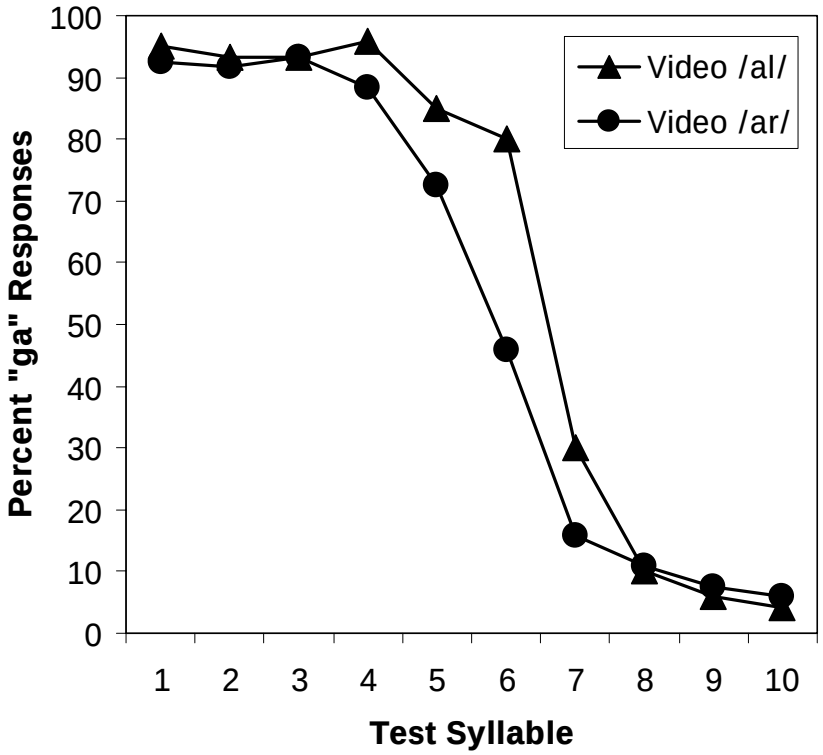

Video /al/
Video /ar/
Percent "ga" Responses
Test Syllable
100
90
80
70
60
50
40
30
20
10
0
1
2
3
4
5
6
7
8
9
10

Figure 7


Video /al/
Video /ar/
Percent "ga" Responses
Test Syllable
100
90
80
70
60
50
40
30
20
10
0
1
2
3
4
5
6
7
8
9
10