Journal of Experimental Psychology:
Human Perception and Performance
1991, Vol. 17, No. 3, 816-828

0096-1523/91/$3.00

Listening With Eye and Hand: Cross-Modal Contributions to Speech Perception

764

Carol A. Fowler
Dartmouth College and Haskins Laboratories,
New Haven, Connecticut

Dawn J. Dekle
Dartmouth College

Three experiments investigated the "McGurk effect" whereby optically specified syllables experienced synchronously with acoustically specified syllables integrate in perception to determine a listener's auditory perceptual experience. Experiments contrasted the cross-modal effect of orthographic on acoustic syllables presumed to be associated in experience and memory with that of haptically experienced and acoustic syllables presumed not to be associated. The latter pairing gave rise to cross-modal influences when Ss were informed that cross-modal syllables were paired independently. Mouthed syllables affected reports of simultaneously heard syllables (and vice versa). These effects were absent when syllables were simultaneously seen (spelled) and heard. The McGurk effect does not arise from association in memory but from conjoint near specification of the same causal source in the environment—in speech, the moving vocal tract producing phonetic gestures.

In a variety of circumstances, including, presumably, face-to-face spoken communications outside the laboratory, listeners can acquire phonetic information optically as well as acoustically. Seeing the face of a speaker considerably improves listeners' abilities to recover speech produced in noise (e.g., Erber, 1969; Ewertsen & Nielsen, 1971; Sumby & Pollack, 1954), and when a visible talker and his or her apparent acoustic output are mismatched in an experiment, as in the so-called "McGurk effect" (e.g., MacDonald & McGurk, 1978; McGurk & MacDonald, 1976), phonetic information recovered optically may override that recovered acoustically. This is particularly likely to occur when the phonetic information is for consonantal places of articulation close to the front of the speaker's mouth. Accordingly, optical *tap* paired with acoustic *map* may be reported as *nap*, with voicing and nasality consistent with the acoustic signal and place of articulation consistent with the optical display (MacDonald & McGurk, 1978; Summerfield, 1987).

Remarkably, the cross-modal influence is not phenomenally due to hearing one utterance, seeing another, and reporting some compromise. Rather, the visible utterance generally changes what the listener experiences hearing (Liberman, 1982; Summerfield, 1987). Accordingly, the visual influence remains when subjects are instructed to report specifically what they heard rather than what the speaker said (e.g., Summerfield & McGrath, 1984), and it remains even after considerable practice attending selectively (Massaro, 1987).

Two general accounts of the McGurk effect may be derived from current theories of speech perception. One is that perceivers consult memory representations of fundamental units of spoken utterances (prototypes of syllables in Massaro's [1987, 1989] theory) that include specifications of both optical and acoustic "cues" for the utterance. Presumably, the memory representations derive from experience with token productions of the utterance in which various subsets of the optical and acoustic cues were detected by the perceiver. When partially conflicting optical and acoustic cues are detected in a McGurk procedure, the listener selects—and experiences hearing—the memory representation most consistent with the collection of cues. In this kind of account, the influence of the visual display on the percept derives from the association of optical and acoustic cues in memory, and these associations, in turn, derive from the association of the cues in the world as sampled by the perceptual systems.

A different general account of the phenomenon derives from theories of speech perception claiming that listeners to speech do not hear the acoustic signal per se but rather hear the phonetically significant gestures of the vocal tract that give rise to the acoustic speech signal. Two such theories are the motor theory of speech perception (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; Liberman & Mattingly, 1985) and the direct-realist theory (Fowler, 1986; Rosenblum, 1987).

The motor theory was developed to explain findings suggesting that at least in some circumstances, there is a closer correspondence between the listener's percept and the vocal-tract gestures of the talker than between the percept and the acoustic signal. (See Liberman et al., 1967, for some examples.) According to the motor theorists, a speech mode of perception in which articulation is recovered from acoustics evolved to handle the special problems that are associated with recovering phonetic information encoded in the acoustic signal. The encoding occurs because consecutive consonants

The research reported here was supported by National Institute for Child and Human Development Grant HD-01994 to Haskins Laboratories.

We thank George L. Wolford for his participation in the pilot research, his guidance on some of the statistical analyses, and for his comments on an earlier draft of the manuscript; we also thank Lawrence Rosenblum for comments on the manuscript and Michael Turvey for comments on our General Discussion section.

Correspondence concerning this article should be addressed to Carol A. Fowler, Department of Psychology, Dartmouth College, Hanover, New Hampshire 03755.

and vowels in an utterance are coarticulated, that is, produced in overlapping time frames. As a consequence of coarticulation, there are no boundaries between phonetic segments in the signal, and the acoustic information for a consonant or vowel is highly context-sensitive. According to the theory, listeners use the acoustic speech signal to devise a hypothesis concerning the set of articulatory gestures that when coarticulated would have given rise to that signal. Testing the hypothesis involves the central component of the listener's speech–motor system (an "innate vocal tract synthesizer" according to Liberman & Mattingly, 1985), and in the theory that is how the percept acquires a motor character.

The direct-realist theory accepts the motor theorists' evidence that listeners to speech recover phonetically significant gestures of the vocal tract, but it explains the recovery differently. In the theory (derived from Gibson's [1966, 1979] more general theory of direct perception), perception is the only means by which organisms can know the environment in which they participate as actors. It is crucial, therefore, that perceptual systems generally acquaint perceivers with relevant aspects of the environment. Perceptual systems do so by extracting information about properties of the environment from media such as light, skin, and air. These media can provide information about the environment, because properties of the environment cause distinctive patternings in them; the distinctive patterns, in turn, can serve as information for their causal source. In vision, although retinas are stimulated by reflected light, perceivers do not see patterns in light; rather, they see the environmental sources of those patterns. In haptics, perceivers do not infer palpated object properties from felt skin deformations; they feel the object sources of the skin deformations themselves. By analogy, in hearing, and specifically in speech perception, listeners should not perceive the acoustic signal but rather its cause in the environment. In speech, the immediate cause of the signal is the vocal-tract activity of the talker.

In either the motor theory or the direct-realist theory, the McGurk effect, and audiovisual influences on speech perception more generally, arise because the optical and acoustic information is convincingly about the same speech event, and speech events, not structured media, are perceptual objects.

In the following experiments, we attempt to distinguish the two general accounts we have offered for the McGurk effect. In its global character, the effect is consistent with both accounts. Perceivers frequently both see and hear a speaker, so they have ample opportunity to develop memory representations including both optical and acoustic cues. By the same token, outside the laboratory, visible talking and audible talking emanating from the same location in space are the same event of talking; if the objects of perception are environmental, as specified by information in media, then it is expected that information in different media that are joint consequences of the same event (or, in the McGurk procedure, of ostensibly the same event) serve jointly to specify the event to the perceiver.

In the following experiments, we have attempted to distinguish the two accounts by looking for cross-modal influences on speech perception from two new sources, each of which captures one but not the other distinctive aspect of the McGurk paradigm that might account for the cross-modal influences there.

On the one hand, one situation was meant to capture the association in experience and hence in memory of an acoustically specified utterance with a specification in another modality. On the other hand, it was meant to exclude association by way of conjoint lawful specification of a common environmental event. To achieve this, we paired spoken syllables with matched or mismatched orthographic representations of syllables (cf. Massaro, Cohen, & Thompson, 1988). Our college-student subjects have been readers of an alphabetic writing system for more than a decade, and they experience redundant pairings of sight and sound whenever they read aloud or else see a text that someone else is reading aloud. Although listeners may be less experienced with sound–spelling pairings than with pairings of the sound and sight of a speaker, their experience with the former pairings are sufficient that a spoken monosyllabic word can activate its spelling for a listener. For example, detection of an auditorily presented word that rhymes with a cue word also presented auditorily is faster if the two words are spelled similarly (e.g., *pie–tie*) than if they are spelled differently (e.g., *rye–tie*) (Seidenberg & Tanenhaus, 1979; see also Tanenhaus, Flanigan, & Seidenberg, 1980). (In an experiment similar to the orthographic condition of our Experiment 1, Massaro, Cohen, & Thompson, 1988, reported weak cross-modal influences of a written syllable on a spoken syllable. Our Experiments 1 and 2 explore the conditions under which this occurs and compare the effect with another possible source of cross-modal influences on speech perception.)

An important characteristic of the acoustic–orthographic pairings for our purposes is that their association is by convention rather than by lawful causation (cf. Campbell, 1989). That is, whereas optical and acoustic correlates of a speaker talking are associated in the world because they are lawful consequences of the same event of talking, graphemes and their pronunciations are associated in the world by societal convention.

The second experimental situation that we devised was meant, insofar as possible, to be complementary to the first. That is, we established a cross-modal pairing that is unfamiliar to subjects but that (outside the laboratory) is a lawful pairing, because the same environmental event gives rise to structure in the two different media. In this situation, we paired acoustically specified syllables with matched and mismatched manually felt, mouthed syllables. Our guess, confirmed by our subjects, was that they did not recollect experiences in which they had handled someone's face while he or she was talking. Although some of them may have had such experiences—as infants or young children perhaps—they must have had considerably less experience of that sort than they had had either seeing and hearing spoken utterances or seeing and hearing text being read.

We know that some phonetic information can be obtained by feeling the face and neck of a speaker. Helen Keller learned to speak English, French, and German by obtaining phonetic information haptically (e.g., Keller, 1903). Other deaf–blind individuals have learned some speech and have learned to understand spoken language remarkably well by using the

Tadoma method, in which they learn to detect phonetic properties of speech haptically (e.g., Chomsky, 1986; Schultz, Norton, Conway-Fithian, & Reed, 1984). We do not know whether inexperienced perceivers can recover phonetic information by touch, and it is important for our test that our subjects not be trained. Accordingly, we selected a fairly distinct articulatory difference for them to detect: the difference between /ba/ and /ga/.

Our expectations were as follows. If the operative factor in the McGurk effect is the association in memory of cross-modal cues available during events in which speech occurs, then an influence of written syllables on heard syllables should occur, but an effect of felt syllables on heard syllables should be weak or absent. If the operative factor is instead the (apparently) common causal source in the environment of the acoustic and optical structure, then felt syllables will affect what listeners report hearing, and orthographic representations of spoken syllables will not. Alternatively, of course, both (or neither) factors may be important.

Experiment 1

Our first experiment tested these predictions by looking for cross-modal haptic and orthographic influences on identifications of heard syllables and for reverse effects of heard syllables on reports of felt and read syllables. In this experiment and the next, we informed subjects that cross-modal syllables were paired independently, and we requested that they therefore make their judgments of heard syllables on the basis of what they heard only; we did so in an effort to restrict cross-modal effects to influences on perception rather than on judgment.

Method

Subjects. Subjects were 23 undergraduate students at Dartmouth College. All were native speakers of English who reported normal hearing and normal or corrected vision. None of the participants was experienced in the Tadoma method, and none had special training in lipreading. Of the 23 participants, one student's data were eliminated from the analyses when she reported in debriefing that she had not believed our (accurate) statement that felt and heard syllables had been paired independently. Data from 10 additional subjects were eliminated from most analyses because their identifications of felt syllables did not significantly exceed chance (60.6% correct; $z = 1.64$).

Stimulus materials. Acoustic stimuli were three-formant synthetic consonant–vowel (CV) syllables produced by the serial resonance synthesizer at Haskins Laboratories. There were 10 syllables in all, ranging across an acoustic continuum of F2 transitions from a rising transition appropriate for /ba/ to a falling transition appropriate for /ga/. F3 was fixed and rising across the continuum; accordingly, there were no intermediate /da/ syllables. Steady-state values for the three formants were 800, 1190, and 2400 Hz (with bandwidths of 50, 100, and 200 Hz, respectively). Starting frequencies of F1 and F3 were 500 and 2100 Hz. Starting frequencies of F2 ranged from 760 Hz at the /b/ end of the continuum to 1660 Hz at the /g/ end in 100-Hz steps. Transitions were 50 ms in duration with a following 150 ms steady state. Fundamental frequency increased by 10 Hz over the first 30 ms of each syllable to a steady-state value of 120 Hz and declined by 10 Hz over the last 40 ms. The amplitude contour had an analogous shape and time course.

The 10 continuum members were stored (filtered at 10 kHz and sampled at 20 kHz) in a small computer (New England Digital Company) programmed to run the various conditions of the experiment. They were presented to listeners over a loudspeaker situated to the right of the cathode-ray tube (CRT) screen.

On each trial in the orthographic condition, "BA" or "GA" was printed on the CRT screen simultaneously with the presentation of the acoustic syllables; the printed syllable remained on the screen until subjects pressed a key to initiate the next trial.

In the Tadoma condition, mouthed syllables were /ba/ and /ga/ produced by a model (Fowler). The model faced a CRT screen that specified the syllable to be mouthed on each trial and provided a countdown permitting the mouthed syllables to be produced in synchrony or near synchrony with the acoustically presented syllable.

A single 60-item test order of the acoustic continuum members was used for all conditions of the experiment. Six tokens of each continuum member were presented in random order but with the constraint that each synthetic syllable occur twice in each third of the test order. In the orthographic condition, a second 60-item test order determined which orthographic syllable would be paired with each acoustic syllable. The same test order dictated the order of syllables to be mouthed in the Tadoma condition. This sequence was also random but now with the constraint that each orthographic (felt) syllable be paired with each synthetic syllable once in each third of the test order.

Procedure. Each subject participated in three tests: synthetic syllables alone (auditory condition), synthetic syllables paired with printed syllables (orthographic), and synthetic syllables paired with felt syllables (Tadoma). The order of the conditions was counterbalanced, with 2 subjects (of the 12 who exceeded chance in identifying felt syllables) experiencing each order.

Subjects were run individually. They first heard the endpoint syllables from the synthetic-speech continuum. The endpoints were identified for them and played several times. Subjects were told that the speech was produced by a computer and that they would be identifying syllables like the ones they had just heard in subsequent tests.

In the auditory test, subjects were seated in front of the CRT screen. Printed on the screen was the message "PRESS RETURN TO PROCEED." To initiate each trial, subjects pressed the return key. On each trial, one synthetic syllable was presented over the loudspeaker. Subjects made their responses by circling printed "B" or "G" on an answer sheet; they were instructed to guess if necessary and then to continue the test at their own pace by pressing the return key for each trial.

In the orthographic condition, the test was similar, except that when the return key was pressed, a printed syllable appeared on the screen with its onset simultaneous with that of the synthetic syllable. The printed syllable remained on the screen until the subject pressed the return key for the next trial.[1] Subjects were instructed to watch the screen as the printed syllable was displayed. They then made two responses, first circling either "B" or "G" under the heading "heard," indicating which syllable they had heard, and then circling "B" or "G" under the heading "saw," indicating which syllable they had seen. Subjects were told explicitly that the acoustic and spelled syllables were independently paired so that they should always base their "heard" judgment on what they had heard independently of

[1] As we discuss later, we attempted to bias the experiment in favor of the orthographic condition by presenting the printed syllable for a long period of time. Possibly, however, this was a design flaw in that the printed syllables failed to demand as much attention as the haptically perceived syllables. We provided a more attention-demanding orthographic condition in Experiment 2.

what they had seen and vice versa for the "saw" judgment. As in the auditory condition, they were instructed to guess if they were unsure of the syllable they had heard or seen on any trial.

In the Tadoma condition, the model stood facing the CRT screen with the loudspeaker directly in front of her at about waist level. She sequenced the trials by pressing the return key to initiate each one. In advance of the acoustic syllable presentation, the computer printed a syllable on the screen that the model was to mouth. Then it presented a countdown consisting of a sequence of two asterisks and then a pair of exclamation points (i.e., *...*... !!) at 1,000-ms intervals. The exclamation points were presented simultaneously with the onset of the acoustic syllable. Between trials, the model kept her lips parted and her jaw slightly lowered. With practice, she learned to time her closing for the mouthed consonant so that the opening coincided phenomenally with the onset of the acoustic signal.

Subjects received no instructions on how to distinguish felt "ba" from "ga." Each subject stood with his or her back to the CRT screen, approximately a step farther from the CRT screen than the model. (This was to prevent the subject from looking at the model's face; in any case, relevant parts of her face were covered by the subject's hand.) The subject stood with his or her right hand in a disposable glove placed over the lips of the model. After some piloting, we found that subjects were most successful if we placed the forefinger on the upper lip and the next finger on the lower lip of the model. This is not the procedure used by Tadoma perceivers (who generally place a thumb vertically against the lips); however, it allowed subjects to distinguish open from closed lips. After presentation of paired felt and heard syllables, subjects indicated to a second experimenter which syllable ("ba" or "ga") they had heard and then which syllable they had felt, in each case guessing if necessary. They made their responses by pointing to printed syllables on an 8½ × 11-in. (21.59 × 27.94 cm) sheet of paper held on a clipboard by a second experimenter (Dekle). On the sheet of paper, the printed syllables "BA" and "GA" appeared twice: on the left under the heading "heard" and on the right under the heading "felt." The second experimenter then marked an answer sheet analogous to those used in the orthographic condition (i.e., with response columns "heard" and "felt"). As in the orthographic condition, subjects were told explicitly that synthetic and mouthed syllables were paired independently and that they should therefore make their "heard" judgment on the basis of what they had heard only and their "felt" judgment on the basis of what they had felt only.

Because the Tadoma procedure was quite taxing, not only for the model but also for subjects (whose right arms were extended to the side of the model's mouth level), subjects were allowed to stop and rest at any point during the procedure. Except for rest periods, trials were sequenced by the model pressing the return key immediately after the second experimenter indicated that she had recorded the subject's responses.

Of course, the model was aware of the syllable to be mouthed but not of the synthetic-syllable continuum member to be presented on each trial. The experimenter who recorded the subject's responses was blind to the syllable being mouthed but able to hear the synthetic syllable. Accordingly, neither experimenter had information that could bias their performance of their respective roles in the experiment with respect to predictions concerning effects of felt speech on heard speech.

Results

Results in the auditory condition for the 12 subjects whose performance in identifying "felt" syllables was significantly better than chance are presented in Figure 1. Performance averaged 90% "b" judgments for the first two continuum members and 12% "b" judgments over the last two. In the cross-modal tasks, performance in identifying the orthographic syllables was, not surprisingly, near 100%, and performance in judging felt syllables averaged 78.6%.

Effects of the orthographic and felt syllables on "heard" judgments are presented in the top and bottom panels of Figure 2. The top panel shows a small effect of the orthographic syllable on the percentage of "b" judgments, and the bottom panel shows a larger, more consistent effect of the felt syllables. In an analysis of variance with repeated measures factors of continuum number, cross-modal syllable (ba/ga), and condition (orthographic, Tadoma), the effect of continuum number on the percentage of "b" judgments was, of course, highly significant, $F(9, 99) = 67.52, p < .001$, accounting for 56.2% of the variance in the analysis. The effect of condition was not significant, but the effect of cross-modal syllable reached significance, as did its interaction with condition, $F(1, 11) = 17.22, p = .0017$; $F(9, 99) = 4.85, p = .048$, respectively.[2] To explore the basis for the interaction, we performed tests separately on the data from each modality, setting alpha now at .025. The 17.5% difference in heard "b" responses accompanied by felt "ba" versus "ga" was highly significant, $F(1, 11) = 13.70, p < .001$, whereas the analogous 5.8% effect in the orthographic condition was marginal, $F(1, 11) = 4.52, p = .055$. In the original analysis, no other interactions reached significance. The main effect of cross-modal syllable and its interaction with condition remained significant, with nearly identical F and p values when the analyses were redone on the percentage of "b" responses averaged across the 10 continuum members.

In the next analysis, we looked only at those "heard" trials on which subjects had correctly identified the "felt" syllables. Because this excluded many trials, we examined performance averaged across the continuum. On average, listeners showed a 30.4% difference in favor of heard "b" judgments for those trials accompanied by felt "ba" rather than felt "ga." This difference is nearly double that found by looking at all heard trials, and it is highly significant: $t(11) = 4.25, p = .0014$. We did not perform an analogous test on the orthographic data, because performance identifying printed syllables was near 100%.

A further analysis was performed on data from the Tadoma condition, now with data from the auditory condition included. If felt "ba" increases heard "b" judgments and felt

[2] In an analysis including all 22 subjects, the main effects of continuum number and of cross-modal syllable remained highly significant, whereas the critical interaction was marginally significant, $F(1, 21) = 4.16, p = .052$. The difference in heard "b" responses as a function of feeling mouthed "ba" versus "ga" was 12.6%; as a function of seeing printed "ba" versus "ga" it was 5.6%. We performed this test including random subjects for two reasons. One was a discomfort with eliminating nearly half of our sample of subjects in the main analyses. A second was a sense that because subjects made the "felt" judgments second, some forgetting might have occurred. If so, their performance on the felt judgments may have underestimated both their true ability to identify felt "ba" and "ga" and the influence that the felt syllable might have had on the heard syllable.

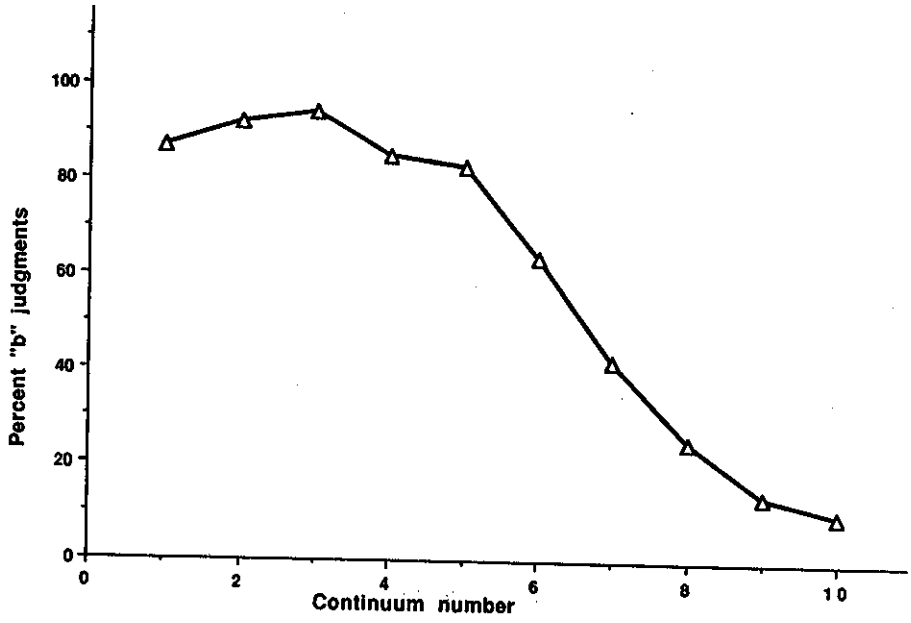

Figure 1. Percentage of "b" judgments to acoustic continuum members in the auditory condition, in which acoustic syllables were unaccompanied by spelled or felt syllables.

"ga" decreases them, the the means for the auditory condition should fall between those for the felt "ba" and "ga" trials. On average, they do, with the percentage of "b" judgments averaging 62.2% overall on felt "ba" trials, 59.6% on acoustic-alone trials, and 44.7% on felt "ga" trials. Although the difference between the acoustic and felt "ga" trials is large and consistent, however, that between the acoustic and felt "ba" trials is not. In an analysis of variance with factors of continuum number and condition (felt "ba," auditory, felt "ga"), both main effects were significant. In analyses performed on each comparison separately, however, only the auditory–"ga" difference was significant, $F(1, 11) = 22.66$, $p = .0006$; the auditory–"ba" difference did not approach significance ($F < 1$). However, the auditory–"ba" variable did interact with continuum number, $F(9, 99) = 2.09$, $p = .037$, with the predicted direction of effect occurring just on Numbers 1, 4, and 7–10. The failure of felt "ba" to differ from the auditory condition may signify that only felt "ga" gave rise to a McGurk-like effect. An alternative interpretation (which we address in Experiment 3) is that blocking the acoustic-alone trials from the Tadoma trials led to a criterion difference in classification judgments across the conditions.

In another analysis on heard syllables in the Tadoma condition, we examined performance on the first block of 20 Tadoma trials to see whether a cross-modal influence was present from the very beginning, when subjects were least experienced with the task. This test is of particular interest in the Tadoma condition, because subjects had essentially no prior experience handling the faces of talkers. We had designed the test order so that every continuum member occurred once with felt "ba" and once with felt "ga" in the first block of 20 trials. This analysis, performed on "b" responses averaged across the continuum, yielded a significant difference between felt "ba" and "ga" trials: $t(11) = 2.84$, $p = .016$. The difference was 17.5% in favor of "b" responses, which coincidentally was a difference of exactly the same magnitude as the difference averaged over all 60 trials.

We looked for a reverse effect of heard syllables on felt judgments. (We did not look for an analogous effect on orthographic judgments, because performance there was at ceiling.) Massaro (1987) obtained an analogous influence of heard syllables on judgments of visually perceived mouthed syllables. In an analysis of variance with factors of continuum number and felt syllable (felt "ba" or "ga"), both main effects and the interaction were significant: continuum, $F(9, 99) = 5.61$, $p < .001$; felt syllable, $F(1, 11) = 61.41$, $p < .001$; interaction, $F(9, 99) = 2.08$, $p = .038$. The main effect of continuum was significant, because there were more felt "ba" judgments on trials associated with acoustic syllables at the /b/ end of the acoustic continuum than on trials associated with acoustic syllables at the /g/ end. The main effect of felt syllable was significant, because there were more felt "ba" judgments when the felt syllable was "ba" (85.9%) than when it was "ga" (29%). To explore the basis for the interaction, we performed separate analyses on the felt "ba" and felt "ga" trials. In both instances, the effect of continuum was significant; we used trend tests to look for a linear effect of acoustic continuum number on the percentage of "ba" judgments. Both tests were significant: felt "ba," $F(1, 99) = 8.54$, $p = .0044$; felt "ga," $F(1, 99) = 29.64$, $p < .001$. In each case, felt "ba" judgments decreased as the acoustic syllable shifted from the /b/ to the /g/ end of the continuum. They decreased from 95% "b" judgments to 80% on felt "ba" trials and from 39% to approximately 10% on felt "ga" trials.

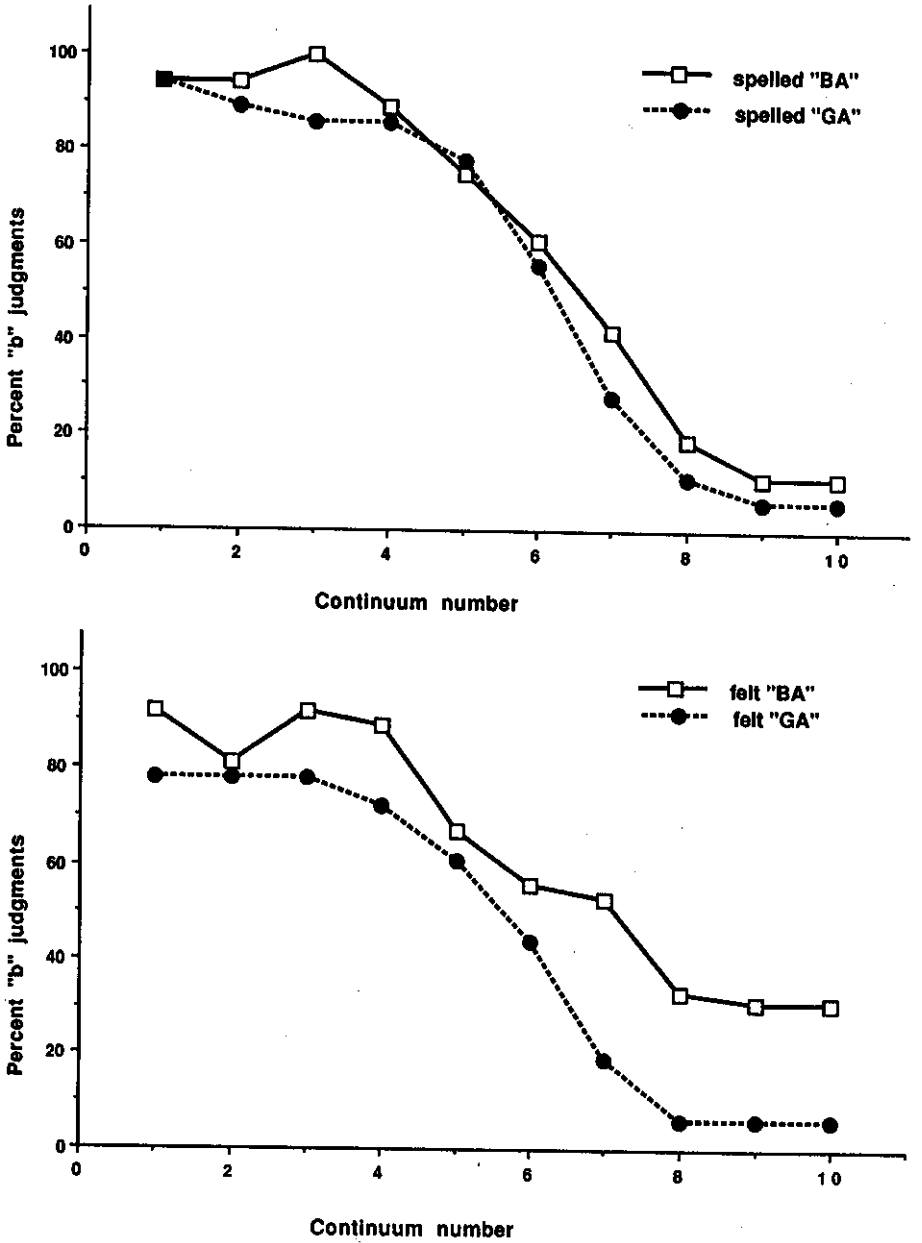

Figure 2. Percentage of "b" judgments to acoustic continuum members in cross-modal conditions of Experiment 1. (Top panel: orthographic condition; bottom panel: Tadoma condition.)

Discussion

The most important outcome of this experiment is that a strong cross-modal effect on judged spoken syllables occurred in one cross-modal condition, whereas at most a marginal effect occurred in the other. In particular, a highly significant effect occurred when mouthed syllables were felt simultaneously with an acoustic presentation of similar syllables, but a marginal or nonsignificant effect occurred when syllables were printed. According to the logic presented in the introduction, we interpret this difference as evidence favoring accounts of speech perception in which perceptual objects are the phonetically significant vocal-tract actions that cause structure in light, air, and on the skin and joints of the hand. By the same token, we have shown that mere association in experience between sources of information for a syllable at best gives rise to a weak cross-modal influence with this procedure. We consider an alternative interpretation of these latter findings, which Experiment 2 is designed to address. First, however, we consider some other outcomes of Experiment 1.

One interesting outcome is that the cross-modal effect of felt syllables on heard syllables is present in inexperienced subjects in the very first block of trials in which they participate. This suggests that the effect arises in the absence or near

absence of experience in which the acoustic and haptic information is paired. Accordingly, we conclude that joint specification of an environmental event does not require specific learning to be effective in perception. We discuss this conclusion further in the General Discussion section.

A second interesting finding of the experiment is that the cross-modal effects in the Tadoma condition worked in both directions. Not only did felt syllables affect judgments of the syllable heard, but the acoustic syllable affected judgments of the syllable felt. This suggests considerable integration of the information from the two modalities.

There is an alternative interpretation of our findings in the orthographic condition. Our expectations had been that the Tadoma condition would lead to cross-modal effects, whereas the orthographic condition would not. Accordingly, we designed the orthographic condition in a way that we thought would optimize its chances of working. That is, we presented the orthographic syllable for a long period of time, guaranteeing that subjects would see the syllable and hence maximizing the number of opportunities for a cross-modal effect to occur. However, the effect of the long-duration presentation may have been different from what we expected. Although we asked subjects to look at the screen as they pressed the return key (so that they would see the printed syllable simultaneously with hearing the acoustic syllable), they need not have followed directions, because the syllable remained on the screen until the next trial was initiated. Furthermore, feeling a mouthed syllable is difficult and attention-demanding in a way that looking at a clearly presented printed syllable is not. One consequence may be that the Tadoma task took attention away from the listening task, leaving the acoustically signaled syllables less clearly perceived and perhaps more vulnerable to influence of information from another modality. To address these possibilities, we designed a second orthographic condition in which printed syllables were presented briefly and masked.

Experiment 2

Experiment 2 was designed to assess the effect of attention on the cross-modal influence of printed syllables on acoustic syllables. We hoped to use masking to force subjects to look at the printed syllable at the same time as they were listening to the acoustic syllable and to drive performance in reporting orthographic syllables down to a level comparable to the nonrandom subjects' ability to report the felt syllables in Experiment 1 (78.6% correct).

Method

Subjects. Subjects were 12 undergraduate students at Dartmouth College who were native speakers of English with normal hearing and normal or corrected vision.

Stimulus materials. Auditory stimuli were those described in Experiment 1. The visual stimuli were also identical to the visual stimuli in Experiment 1, except they were masked by a row of number signs ("#"). On each trial, a row of four number signs appeared just above the location on the screen where the syllable would be printed. Simultaneous with presentation of the synthetic-speech syllable, either "BA" or "GA" was printed on the screen below the number signs; the printed syllable was covered after 67 ms by another row of number signs.[3] The mask remained on the screen until the subject hit the key for the next trial.

The 60-trial test orders of continuum members and orthographic syllables used in Experiment 1 were also used here to sequence stimulus presentations.

Procedure. Each subject was tested individually. As in Experiment 1, the endpoints of the continuum were played and identified before the main test to allow subjects to become familiar with the synthetic-speech syllables. Subjects were told that they could pace themselves through the experiment by hitting the return key, and they were instructed not to leave any blanks on the answer sheet, taking a guess if necessary. In addition, as in Experiment 1, subjects were told that acoustic signals and masked visual signals were independently paired on each trial, and therefore decisions about them should be made independently.

Results and Discussion

We succeeded in lowering performance on the printed syllables to the level of the nonrandom subjects in the Tadoma condition of Experiment 1. Performance in judging orthographic syllables averaged 78.3%, nearly the same level of performance in judging felt syllables found in the Tadoma condition.

The major results of the experiment are depicted in Figure 3. The figure shows no effect of the orthographic syllable on identifications of heard syllables. An analysis of variance on the data in Figure 3 revealed only a significant effect of continuum number: $F(9, 99) = 39.87$, $p < .001$. The effect of orthographic syllable was nonsignificant, with a 4.4% numerical difference between means in the unpredicted direction; the interaction also failed to reach significance. In an analysis of variance including the Tadoma condition of Experiment 1 and the results of the present experiment, the Condition × Cross-Modal Syllable interaction was highly significant, $F(1, 22) = 17.18$, $p = .0005$.

An analysis of variance on the effect of the acoustic signal on visual judgments yielded two significant effects. The effect of orthographic BA or GA was, of course, highly significant, $F(1, 11) = 319.02$, $p < .001$, accounting for 60.6% of the variance. In addition, there was a marginal effect of acoustic continuum number, $F(9, 99) = 1.98$, $p = .049$, which accounted for just 1.4% of the variance. The interaction between continuum and orthographic BA or GA did not reach significance. We performed a trend test on the effect of continuum number to determine whether decreases in identifications of an orthographic syllable such as BA were associated with decreasingly /ba/-like acoustic signals. The analysis, which tested for a linear decrease in "ba" identifications across the continuum, did not approach significance ($F < 1$).

[3] The stimulus onset asynchrony (SOA) was determined by pilot-testing subjects until we found an SOA at which performance in identifying the spelled syllables was approximately that of the haptically identified syllables of Experiment 1. The SOA at which performance approached that of the Tadoma condition of Experiment 1 was 67 ms. Occasional trials may have had SOAs of 83 ms, however, because we were unable to control target and mask presentations in relation to the occurrence of screen refreshes.

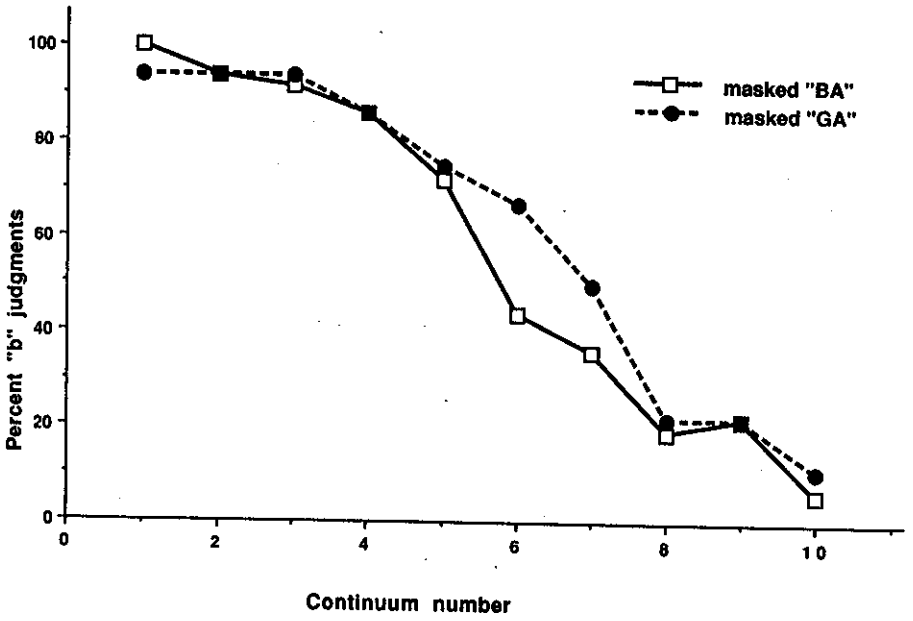

Figure 3. Percentage of "b" judgments to acoustic continuum members in the masked orthographic condition of Experiment 2.

We were successful in using masking to decrease identifiability of the orthographic syllables to a level comparable to identifiability of felt syllables among nonrandom subjects in the Tadoma condition. We hoped that this manipulation would increase attention to the visual stimulus while the acoustic signal was being presented and would attract attention away from the acoustically presented syllables. If these attentional features had been characteristic of the Tadoma condition of Experiment 1 but not of the orthographic condition of that experiment, and if those differences had been the operative ones in the different outcomes of those conditions in Experiment 1, then a cross-modal effect should have been strengthened in the present experiment. Instead, the marginal effect in Experiment 1 disappeared completely in Experiment 2; accordingly, we revert to our original interpretation of the difference in outcome. McGurk-like effects occur when information from the two modalities are conjoint, lawful consequences of the same environmental event. They do not occur only on the basis of association in experience.

Experiment 3

We performed a final experiment in two phases to address a variety of methodological concerns raised by reviewers about the haptic condition of Experiment 1. Two concerns were consequences of our having used a loudspeaker rather than headphones to present the acoustic stimuli. Possibly mouthed syllables were whispered, and the influence on heard syllables in that condition was auditory–auditory, not haptic–auditory. In addition, because the model could hear the spoken syllables, perhaps that biased her mouthing actions. Although we discount both concerns, it was easy to allay them by eliminating the loudspeaker and substituting headphones.[4] Of course, this is likely to reduce the cross-modal influencing effect, as do other manipulations that decrease the compellingness of information that the cross-modal signals derive from the same physical event (e.g., see Easton & Basala, 1982). However, we considered our original effect large enough to survive some reduction because of spatial dislocation of voice and mouth.

A third concern was that we had provided no measure of synchrony between mouthed and audible syllables; we provided a measure, albeit a phenomenal one, in Experiment 3. A fourth concern was with the dropout rate among our subjects. We believed that the poor performance levels on haptic judgments on the part of some subjects in Experiment 1 were due to discomfort with the procedure and the absence of rewards for high performance. Accordingly, in Experiment 3 we instituted rewards.

[4] In Experiment 1, only supraglottal actions of the model mimicked those of naturally spoken /ba/ and /ga/; there was no channeling of air through the oral cavity and, accordingly, there was no whispering. As for biasing effects on the model of hearing the synthetic syllables, two factors precluded that. First, to make the movements (particularly of /ba/) synchronous with the synthetic syllable—a condition we presumed important to integration (cf. Easton & Basala, 1982)—closing movements for the consonant had to be initiated before the onset of the signal so that articulatory release, where the onset of acoustic energy begins for stop consonants, co-occurred with the stop burst and vowel onset. Second, however, if the model were to delay mouthing the syllable until the identity of the spoken syllable was detectable, no behavior other than following instructions to mouth the syllable printed on the CRT screen could have led subjects to shift their judgments of the heard syllable in the direction of the mouthed syllable.

The most serious concern with the findings, however, was that they are consistent with two interpretations of the cross-modal influence of felt mouthed syllables on heard syllables. One is that the influence reflects a true integration, within a trial, of information from two modalities ostensibly about the same physical event. The other is that information from the two modalities remains unintegrated and that over trials, subjects sometimes select either the acoustic syllable or the haptic syllable and give it as their response both to the heard and to the felt syllable. We had attempted to discourage such a response strategy by informing our subjects that we had independently and randomly paired mouthed and spoken syllables and that in consequence, the identification of a syllable in one modality provided no information about the identification of the syllable in the other modality. Moreover, the selection strategy was equally available to subjects in Experiment 2, but it was not adopted there. We interpreted this as evidence that subjects understood, believed, and could follow those instructions. Accordingly, we ascribed the cross-modal influences that occurred in the haptic condition but not the masked visual condition as evidence of integration taking place in one set of conditions and not in the other. In Experiment 3, however, we attempted to provide more direct evidence for integration rather than selection.

Massaro (1987) provided two kinds of information for distinguishing integration and selection in his tests of the effects of visible speaking on identification of spoken syllables. One kind of information is provided by superadditivity of contributions from the modalities to response identifications. In one of Massaro's experiments, on average, subjects correctly identified visible /ba/ presented with no accompanying acoustic signal on 75% of trials. Furthermore, they identified, for example, the most /ba/-like acoustic syllable presented unimodally as "ba" on slightly more than 80% of trials. However, the same acoustic continuum member accompanied by visible /ba/ was identified as "ba" nearly 100% of the time. When that pattern is present to a sufficient degree in individual-subject performances, it rules out one version of a selection strategy, because selection by that strategy will yield an averaging of identification percentages weighted by the relative frequencies with which each modality is selected. The strategy is available to subjects who can process just one modality of information on each trial. In that case, they have just one response to offer, and they may give that response as their identification of syllables presented in both modalities. That strategy will mimic evidence of cross-modal integration of information, except that the probability of identifying a given acoustic syllable, say, as "ba" on a bimodal trial, cannot exceed a weighted average of the probabilities of identifying each unimodal syllable as "ba." In the aforementioned example, performance identifying spoken /ba/ to the first acoustic continuum member could not exceed 80% on bimodal trials, given that the acoustic syllable was identified as "ba" 80% of the time, and the visible syllable was identified as "ba" less than 80% of the time on unimodal trials.

We looked for evidence of superadditivity in our data, but we point out here that it does not provide strong evidence against selection.[5] It is implausible to assume that subjects can only process information from one modality at a time. More likely, both modalities provide perceptual information on each trial. A possible selection strategy, then, is, for example, in identifying the heard syllable, to select a response on the basis of auditory information (so, in the example, identify the syllable as "ba" with probability .8) unless that information is noisy; in that case, use the information from the other modality (in the example, choose "ba" with probability .75). Now the bimodal probability becomes $.8 + (1 - .8).75 = .95$, a value higher than either unimodal value.

A second kind of information for integration rather than selection is provided by evidence that the syllables from the modalities blend in some way. For example, given visible /da/ and auditory /ma/, an identification of the spoken syllable as "na" blends place information from the visible syllable and manner information from the acoustic syllable. In Experiment 3, we also looked for evidence of blending.

Method

Subjects. Subjects were 10 undergraduate students at Dartmouth College who participated for course credit. Two of the 10 subjects were also paid for high performance in identifying haptic and continuum endpoint acoustic syllables. All subjects reported normal hearing. Data from 3 unpaid subjects were eliminated because their performance in identifying haptic syllables was at chance.

Stimulus materials. Acoustic stimuli were those used in Experiments 1 and 2. We created four test orders of 96 trials each. Of the 96 trials, 60 were cross-modal audiohaptic syllables as in Experiment 1. In 30 trials, the 10 acoustic continuum members were presented by themselves three times each. Six trials were unimodal haptic trials in which the mouthed syllables BA and GA occurred three times each. Unimodal trials are needed to test for superadditivity of bimodal contributions to syllable identification. Trials of the various types were randomized in each test order.

Procedure. Subjects participated in four sessions; in each of which, one of the test orders was used. The order of the different test sequences was varied across subjects. Generally, instructions were the same as in Experiment 1, except of course that we told subjects that some trials were not bimodal and that they should just report one syllable on those trials. An additional change in procedure was that subjects now wore headphones over which acoustic syllables were played; the acoustic syllables were thereby inaudible to the experimenters. A final change in procedure, meant to improve performance in identifying haptic syllables, is that we invited subjects to use whatever hand placement on the face made it easiest for them to identify mouthed syllables.

For the first 2 subjects we ran, we offered a system of payments for high performance on haptic identifications and on identifications of acoustic continuum endpoints. In particular, subjects could receive \$0.25 for each percentage point above 75% for haptic identifications and another \$0.25 for each percentage point above 75% for identification of acoustic endpoints. Subjects could therefore earn a maximum of \$50 in all across the four sessions.

We also asked these same subjects to give us a final judgment on each trial in addition to identification of heard and/or felt syllables. We asked them to tell us whether the syllables in the two modalities were synchronous. In particular, we told them that they should report "E" if the mouthed syllable led the acoustic syllable, an "S" if it was synchronous, and an "L" if the mouthed syllable lagged the acoustic syllable.

---

[5] We thank George Wolford for pointing this out.

Both of these latter procedures were abandoned after 2 subjects were run. The system of payments was abandoned because performance on haptic unimodal syllables was near perfect, making it impossible to test for integration by looking for superadditivity. (One subject earned $43 of the maximum possible of $50; the other earned $39.50. On unimodal haptic trials, performance averaged 92% correct—too high to look for superadditivity of cross-modal influences on syllable identification.) The assessment of simultaneity was abandoned because even though we did not bias subjects to expect simultaneity, they reported simultaneity virtually all of the time (on 97% of trials for 1 subject and 98% for the other).

Because we feared that superadditivity would remain difficult to test for and because we have learned that it does not provide a strong test of integration anyway, we made a final change in procedure that we hoped would enable us to test for integration by testing for blend responses. For the remaining subjects, we opened the response inventory in the following way. As we had done in Experiments 1 and 2 and in running the first 2 subjects of Experiment 3, we played the acoustic continuum endpoints to subjects before they began the experimental test. In addition, we told the remaining subjects that there were 10 syllables in all and that the other 8 ranged between the /ba/ and /ga/ syllables they had just heard. We played them Continuum Member 5 as an example of an ambiguous syllable. We told subjects that we did not know how listeners would identify those intermediate sounds; they might sound like ambiguous /ba/s and /ga/s, or they might sound like other consonants to them, possibly /ða/, /va/, or /da/. We asked them always to report the sound they heard, whether it was /ba/, /ga/, or some other consonant sound. We also gave them to understand that, haptically, although they should feel some clear instances of mouthed /ba/s and /ga/s, sometimes they might feel other consonant-initial syllables. On judgments of felt syllables, they should always report the syllable they experienced, whether it was /ba/, /ga/, or some other consonant-initial syllable.

Because our procedural changes made little difference in subjects' performances, for most analyses, we have pooled the findings on all subjects who completed the four sessions. (Three of the 8 unpaid subjects were dropped after one or two sessions for chance performance on haptic trials. Thus, after abandoning our system of payments, performance that had been at ceiling on unimodal haptic trials and was very high on bimodal trials reverted to a level just 3% higher than its level in Experiment 1. We believe that subjects can do the haptic task if they are motivated to attend.)

Results

Figure 4 presents the judgments of heard syllables on the auditory-only (top) and cross-modal (bottom) trials of Experiment 3. In an analysis of variance on cross-modal trials with factors of continuum number and haptic syllable, both main effects were significant: continuum, $F(9, 54) = 74.60$, $p < .0001$; haptic syllable, $F(1, 6) = 16.29$, $p = .007$. The interaction did not approach significance ($F < 1$). Although the effect of haptic syllable was numerically smaller than in Experiment 1, perhaps because of the spatial dislocation of cross-modal syllables, it is present on nearly every syllable along the continuum and is highly significant.

Another reason for the reduced cross-modal effect as compared with its magnitude in Experiment 1 might be that subjects learned over sessions to divide their attention across sessions. To look for effects of practice, a further test was performed, now on the percentage of "ba" responses averaged across the continuum and now including auditory-alone trials, with session as an independent variable. In the analysis, there was an effect of haptic syllable—BA, GA, none, $F(2, 12) = 4.81$, $p = .029$—and an effect of block—$F(3, 18) = 3.20$, $p = .048$—however, the interaction did not approach significance ($F < 1$). As in Experiment 1, "ba" responses to auditory-alone syllables (49.5%) fell between responses to syllables accompanied by mouthed BA (54%) and responses to syllables accompanied by mouthed GA (49%). Now, however, they fell closest to (and very close to) performance on mouthed GA trials rather than mouthed BA trials as they had in Experiment 1. The effect of sessions occurred because the percentage of "ba" responses declined monotonically from Session 1 (54.8%) to Session 4 (48.7%). However, the size of the cross-modal effect did not change across sessions, and the numerical change was in the direction of an increased, not a decreased, effect with practice.

We looked also at the effect of acoustic syllables on haptic judgments. We compared the percentage of "ba" judgments with mouthed BA and GA when they were accompanied by acoustic syllables from the /ba/ end (Items 1–5) or the /ga/ end items (6–10) of the continuum. Table 1 gives the results. For both mouthed syllables, identification of the felt syllable as "ba" was more likely when the mouthed syllable was produced synchronously with a /ba/-like acoustic syllable than when it was produced with a more /ga/-like syllable. An analysis of variance was performed on the data in Table 1, with arcsine transformed to eliminate variance differences due to approaches to ceiling for mouthed BA at the /ba/ end of the continuum and to floor of mouthed GA produced at the /ga/ end of the continuum. The effect of haptic syllable was, of course, highly significant, $F(1, 6) = 45.69$, $p = .0007$. The effect of acoustic syllable was also significant, $F(1, 6) = 6.42$, $p = .04$, as was the interaction, $F(1, 6) = 9.72$, $p = .02$. The interaction occurred because the effect of continuum was 28.7% for haptic GA but only 7.7% for haptic BA, where performance was very close to ceiling.

We turn now to tests of integration. Generally, the test for superadditivity was not applicable to subjects, because their performance in identifying haptic syllables was at ceiling on unimodal trials. In particular, none of our 7 subjects was less than 92% correct in identifying haptic BA as such on haptic-alone trials; 5 subjects made no errors. Obviously, we cannot look for superadditivity of effects of haptic and acoustic information with haptic- and acoustic-alone trials to predict performance on bimodal trials. We do not know why Massaro's subjects performed relatively poorly (75% accurate in his Figure 7; 1987, p. 65) in identifying /ba/ in a video-only condition so that evidence of superadditivity could be sought. By the same token, 3 of our 7 subjects made no more than 8% errors in identifying /ga/ on haptic-alone trials. We looked at performances of the remaining 4 subjects. Of them, 2 showed some evidence of superadditivity. Across haptic and acoustic identifications on trials involving GA as a mouthed syllable, there are 20 opportunities to show superadditivity of haptic-alone and acoustic-alone contributions to bimodal identification of the mouthed or acoustic syllables. Two of the 4 subjects showed superadditivity on 6 of the 20 trials; of the remaining 2 subjects, 1 showed superadditivity on two trials and 1 on none. We do not know how to evaluate this

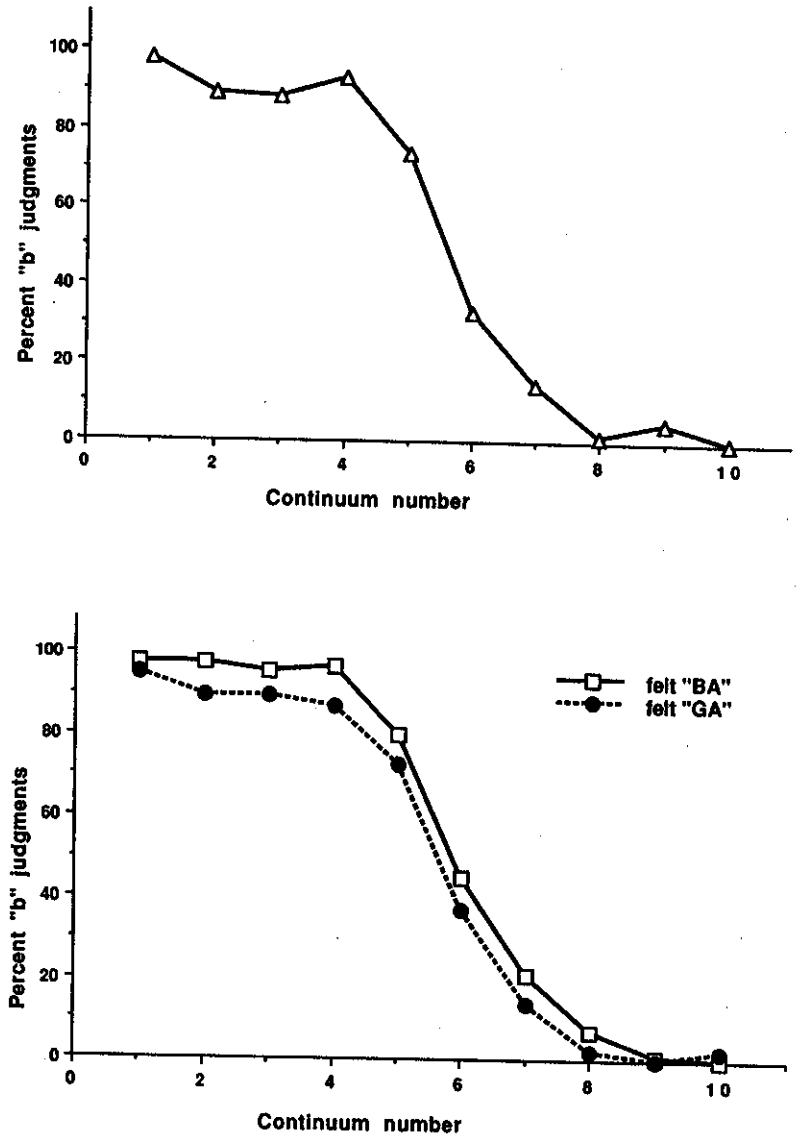

Figure 4. Percentage of "b" judgments to acoustic continuum members on the auditory-alone trials (top panel) and cross-modal trials (bottom panel) of Experiment 3.

outcome statistically. (In Massaro's grouped data [1987, p. 65] four of nine continuum members show superadditivity in relation to video-alone /ba/, whereas none do so in relation to video-alone /da/, where performance is at ceiling; in his sample individual subject data, two continuum members show superadditivity in relation to video-alone /ba/, whereas none do in relation to video-alone /da/, where performance is again at ceiling. Massaro evaluated superadditivity by testing his model, in which bimodal influences are integrative and hence superadditive against a competing selection model. Because we reject Massaro's model as a source of cross-modal haptic influences on auditory speech perception and vice versa, we did not copy his procedure.)

A second test for integration involves tests for blends of information presented in the different perceptual modalities. To enable us to test for blends, we opened the response inventory for subjects after the first 2 subjects that we tested in Experiment 3. However, just 1 subject of the remaining 5 used a response other than "b" or "g" on more than a dozen occasions. The 1 subject used "va" as a frequent response for both heard- and felt-syllable judgments. Here we consider the patterning of his "va" responses.

Because /ba/ and /ga/ are articulatory extremes in English, almost any consonant response other than "b" or "g" is intermediate between /ba/ and /ga/ and hence looks, on the surface, like a blend. To attempt to determine whether the subject who gave frequent "va" responses was giving true blends or, alternatively, was simply guessing by giving another consonant name, we looked for a systematic patterning in "va" responses. We reasoned that if "va" responses are blends, then when BA is the mouthed syllable, "va"s should be given as an indication of either the heard syllable or the felt syllable

Table 1
Percentage of "ba" Responses to Haptic BA and GA on Trials in Which Mouthed Syllables Were Produced Simultaneously With Synthetic Syllables From the /ba/ End (Continuum Numbers 1–5) and /ga/ End (Continuum Numbers 6–10) of the Continuum

Syllable	/ba/ end	/ga/end
Haptic BA	95.4	87.6
Haptic GA	35.7	7.0

more frequently when acoustic stimuli are at the /ga/ end of the continuum than when they are at the /ba/ end. That is, as the haptic syllable pulls the response /ba/-ward, the acoustic syllable pulls it /ga/-ward, yielding a blend response. Similarly, when GA is the mouthed syllable, "va" identifications should be more frequent when acoustic syllables are at the /ba/ end than at the /ga/ end of the continuum. For purposes of analysis, we eliminated Middle Continuum Syllables 5 and 6 from the analysis and examined identifications of acoustic and haptic syllables when acoustic syllables were /ba/-like (Continuum Numbers 1–4) or /ga/-like (Continuum Numbers 7–10). Table 2 shows the results for the single subject who gave a considerable number of "va" responses.

His judgments of both acoustic and haptic syllables show the predicted pattern. That is, when BA is the mouthed syllable, "va" identifications of both the acoustic syllables and the haptic syllables are more likely when the acoustic syllable is from the /ga/ end than from the /ba/ end of the continuum. The pattern is reversed when the mouthed syllable is GA. Neither has sufficiently high frequencies for a chi-square analysis; however, a pooled table does. In an analysis of the likelihood of a "va" identification of either a haptic or acoustic syllable when haptic and acoustic syllables pull in the same or opposite directions, the chi-square value is highly significant: $\chi^2(1) = 8.54$, $p = .0036$. On the basis of this analysis, we can conclude that there exists at least one person for whom the influence of haptic judgments on acoustic judgments and vice versa is a true integration and not a selection. Possibly, 2 of the other 7 subjects show evidence for integration in the form of superadditivity of cross-modal influences as well.

Table 2
One Subject's Frequency of "va" Responses to Acoustic (Left) and Haptic (Right) Syllables on Bimodal Trials, When the Synthetic Syllables Were From the /ba/ or the /ga/ Ends of the Continuum

	Heard syllable		Felt syllable	
Syllable	/ba/ end	/ga/ end	/ba/ end	/ga/ end
Haptic BA	5	6	3	5
Haptic GA	10	1	14	4

Note. Middle Continuum Numbers 5 and 6 were disregarded.

General Discussion

Our findings in the Tadoma condition of Experiment 1 and in Experiment 3 suggest that perception need not be based on covert anticipations of the range of sensory cues that may be experienced from stimulation or on associations between those cues and representations of their environmental causes. Together with the results of the orthographic conditions of our experiments, the findings suggest indeed that stored associations are not sufficient for cross-modal integration of information for an event, and that association in the world, specifically due to (ostensible) joint causation by a common environmental event, is required.

In our view, there is another indication that perception of speech syllables does not require prior existence in memory of a prototype or some other way of associating sensory cues to representations of perceivable events. In a McGurk procedure, when the acoustic syllable is /da/ while the model mouths /ba/, listeners frequently report having heard /bda/ (e.g., Massaro, 1987). In Massaro's fuzzy-logical model of perception, a syllable is reported by listeners if it is represented by a prototype in memory and if evidence consistent with that prototype is stronger than evidence consistent with other prototypes as determined by Luce's (1959) choice rule. In discussing the definition of prototypes in the model, Massaro referred to a prototype for /bda/ (1987, p. 128). But how could a hearer acquire a prototype for /bda/, which is not a possible syllable in English? Indeed, it violates the language-universal sonority constraint (roughly that consonants in a within-syllable cluster must increase in vowel-likeness toward the syllable's vowel; e.g., see Selkirk, 1982). Listeners will not have experienced /bda/ prior to the experiment, nor will evolution have provided a prototype for a universally disallowed syllable.[6] The only possibility, it seems, is that prototypes can be constructed "on the fly" in the experiment. That is, there must be a way for the perceiver to decide that no existing prototype is sufficiently supported by the evidence in stimulation. In that case, a new prototype is established and named. However, if the information in stimulation is sufficient to reveal that a new prototype is needed and to name the prototype /bda/, then it seems that the prototype itself is not needed for perception. Rather, it depends on perception for its establishment.

We also doubt that our outcome is compatible with the motor theory. Liberman and Mattingly (1985) invoke an innate vocal-tract synthesizer that subserves both speech production and, through the use of something like analysis by synthesis, speech perception as well. There is reason to sup-

[6] Although in rapid speech production certain "be-" words (such as "beneath" or "become") may be reduced so that the first vowel is inaudible, there are no entries in the dictionary whereby such reduction would lead to a /bda/ syllable. (Readers of the manuscript have suggested the words "bidet," "bedazzle," "bedecked," and "bedevil"; however, they should recall that Massaro's memory representations are syllable prototypes, not consonant prototypes. Were the schwa vowels dropped in the foregoing words, the remaining syllables would be /bdey/, /bdaez/, /bdɛk/, and /bdɛv/, not /bda/ (/a/ being the first vowel in "father").

pose that an innate synthesizer would have anticipated the possibility of receiving optical as well as acoustic information about vocal-tract gestures because both of those information sources are commonly available in the environment of listeners; ability to exploit the information sources might have adaptive significance. However, there is no reason to suppose that selection would have favored evolution of a synthesizer that anticipated receiving haptic information provided by the hands of a listener. More generally, we doubt that any explanation for the Tadoma effect can work that depends on the significance of the haptic information being appreciated either because the significance has been learned or because it is known innately.

How can a new environmental occurrence be perceived, or likewise in our experimental situation, how can a familiar occurrence be signaled effectively, in part by novel proximal stimuli? We follow Shaw, Turvey, and Mace (1982) in concluding that perceptual experience is fundamentally knowledge acquired because of the "force of existence" of events in the world. In our terms (and not necessarily those of Shaw et al.), environmental events causally structure media such as light, air, and the skin and joints of a perceiver. To the extent that the structure is distinctive to its causal source, it can serve as information about the source. The information can inform without prior familiarity with it because of the causal chain that supports perception. Stimulation caused by an environmental event has causal effects on sensory receptors so that its structure is, in part, transmitted to a perceptual system. By hypothesis, the perceiver comes to know an event in the environment by way of its impact on the perceptual systems as transmitted by proximal stimulation.

In that sense, perception may not be a particularly intellectual endeavor at all; it may be more analogous to motor accommodations to or motor exploitations of physical forces exerted on the bodies of actors. In the analogy, proximal stimuli at the sense organs are the informational analogs of the physical forces impinging on the body of an actor. We need not learn what haptic consequences of environmental events mean to perceive their source any more than we have to learn what gravity means to be affected by it in a coherent way. The role of learning, then, may be to change the perceiver–actor's state of preparedness for receiving and exploiting the "forces," both physical and informational, that the world has to offer, not to discover what environmental events the forces signal. That is given in the stimulation. Even in the absence of relevant learning by the perceiver–actor, the environment exerts its same forces that causally affect the body, including the perceptual systems.

References

Campbell, R. (1989). Seeing speech is special. *Behavioral and Brain Sciences, 12,* 758–759.

Chomsky, C. (1986). Analytic study of the Tadoma method: Language abilities of three deaf-blind subjects. *Journal of Speech and Hearing Research, 29,* 332–347.

Easton, R., & Basala, M. (1982). Perceptual dominance during lipreading. *Perception & Psychophysics, 32,* 562–570.

Erber, N. (1969). Interaction of audition and vision in the recognition of oral speech stimuli. *Journal of Speech and Hearing Research, 12,* 423–425.

Ewertsen, H., & Nielsen, H. B. (1971). A comparative analysis of the audiovisual, auditive and visual perception of speech. *Acta Otolaryngolica, 72,* 201–205.

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

Gibson, J. J. (1966). *The senses considered as perceptual systems.* Boston: Houghton Mifflin.

Gibson, J. J. (1979). *The ecological approach to visual perception.* Boston: Houghton Mifflin.

Keller, H. (1903). *The story of my life.* New York: Doubleday.

Liberman, A. (1982). On finding that speech is special. *American Psychologist, 37,* 148–167.

Liberman, A., Cooper, F., Shankweiler, D., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review, 74,* 431–461.

Liberman, A., & Mattingly, I. (1985). The motor theory revised. *Cognition, 21,* 1–36.

Luce, D. (1959). *Individual choice behavior.* New York: Wiley.

MacDonald, J., & McGurk, H. (1978). Visual influences on speech perception processes. *Perception & Psychophysics, 24,* 253–257.

Massaro, D. (1987). *Speech perception by ear and eye: A paradigm for psychological inquiry.* Hillsdale, NJ: Erlbaum.

Massaro, D. (1989). [Review of *Speech perception by ear and eye: A paradigm for psychological inquiry*]. Behavioral and Brain Sciences, 12, 741–755.

Massaro, D., Cohen, M., & Thompson, L. (1988). Visible language in speech perception: Lipreading and reading. *Visible Language, 22,* 8–31.

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

Rosenblum, L. (1987). Towards an ecological alternative to the motor theory of speech perception. *Perceiving–Acting Workshop Review, 2,* 25–28.

Schultz, M., Norton, S., Conway-Fithian, S., & Reed, C. (1984). A survey of the use of the Tadoma method in the United States and Canada. *Volta Review, 86,* 282–292.

Seidenberg, M., & Tanenhaus, M. (1979). Orthographic effects on rhyme monitoring. *Journal of Experimental Psychology: Human Learning and Memory, 5,* 546–554.

Selkirk, E. (1982). The syllable. In H. van der Hulst & N. Smith (Eds.), *The structure of phonological representations, Vol. 2* (pp. 337–384). Dordrecht, The Netherlands: Foris.

Shaw, R., Turvey, M. T., & Mace, W. (1982). Ecological psychology: The consequence of a commitment to realism. In W. Weimer & D. Palermo (Eds.), *Cognition and the symbolic processes, 2* (pp. 159–226). Hillsdale, NJ: Erlbaum.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America, 26,* 212–215.

Summerfield, A. Q. (1987). Some preliminaries to a comprehensive account of audiovisual speech perception. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip-reading* (pp. 3–51). London: Erlbaum.

Summerfield, A. Q., & McGrath, M. (1984). Detection and resolution of audio-visual incompatibility in the perception of vowels. *Quarterly Journal of Experimental Psychology, 36A,* 51–74.

Tanenhaus, M., Flanigan, H., & Seidenberg, M. (1980). Orthographic and phonological activation in auditory and visual word recognition. *Memory & Cognition, 8,* 513–520.

Received April 20, 1990
Revision received October 29, 1990
Accepted October 30, 1990 ■