*Perception & Psychophysics*
*1992, 51 (4), 355-362*

# Audiovisual integration in perception of real words

DAWN J. DEKLE
*Dartmouth College, Hanover, New Hampshire*

CAROL A. FOWLER
*Dartmouth College, Hanover, New Hampshire*
*and Haskins Laboratories, New Haven, Connecticut*

and

MARGARET G. FUNNELL
*Dartmouth College, Hanover, New Hampshire*

Three experiments follow up on Easton and Basala's (1982) report that the "McGurk effect" (an influence of a visibly mouthed utterance on a dubbed acoustic one) does not occur when utterances are real words rather than nonsense syllables. In contrast, with real-word stimuli, Easton and Basala report a strong reverse effect whereby a dubbed soundtrack strongly affects identification of lipread words. In Experiment 1, we showed that a strong McGurk effect does obtain when dubbed real words are discrepant with observed words in consonantal place of articulation. A second experiment obtained only a weak reverse effect of dubbed words on judgments of lipread words. A final experiment was designed to provide a sensitive test of effects of lipread words on judgments of heard words and of heard words on judgments of lipread words. The findings reinforced those of the first two experiments that both effects occur, but, with place-of-articulation information discrepant across the modalities, the McGurk effect is strong and the reverse effect weak.

In the "McGurk effect," first observed by McGurk and MacDonald (1976; see also MacDonald & McGurk, 1978), a listener's perceptual report of a heard nonsense syllable can be influenced by the sight of a speaker producing a different syllable. In one experiment, MacDonald and McGurk dubbed a videotape in which a speaker produced various consonant-vowel (CV) syllables. On the dubbed tape, the mouthed syllable and the syllable on the soundtrack were discrepant. Syllables that subjects reported hearing were influenced by the mouthed syllable. For example, mouthed /da/ paired with acoustic /ma/ was most frequently identified as "na." The manner-place hypothesis put forth by MacDonald and McGurk seems to account for much of the data (but see Summerfield, 1987, for some qualifications). The hypothesis suggests that the acoustic signal dominates in perception of manner of articulation (and voicing) where optical information is poor or absent. In contrast, optical information dominates in perception of those places of articulation that are visibly apparent.

Although an effect of visual information was not anticipated by any theory of speech perception, a natural accounting of it is available in at least three current theories. In two, the motor theory (e.g., Liberman & Mattingly, 1985) and the direct-realist theory (Fowler, 1986; Rosenblum, 1989), listeners to speech are claimed to perceive linguistically significant actions of the vocal tract, which are signaled to different degrees both optically and acoustically. In the fuzzy-logical theory of speech perception (Massaro, 1987), any cue associated with production of a syllable in experience, and therefore in memory, can serve as information for the syllable.

Our purpose in the present series of experiments is not to distinguish among the theories (but see Fowler & Dekle, 1991); rather, it is to follow up on a set of findings reported by Easton and Basala (1982) that are unexpected from all three theoretical perspectives. In a comparison of visual influences on reports of heard words and of acoustic influences on reports of mouthed words, Easton and Basala found no visual influence—that is, no McGurk effect—but a strong reverse influence of acoustic signals for words on judgments of mouthed words. They ascribed their failure to obtain a McGurk effect to their use of real words, which provided "additional speech information ... especially transitions to and from phonemes" (p. 570). This, of course, suggests that visual influences are largely absent in speech perception outside the laboratory, where phonemes are almost invariably embedded in real words. The accounts of the McGurk effect offered by the motor theory, the direct-realist theory, and the fuzzy-logical theory agree in failing to predict Easton and Basala's out-

This research was supported in part by NICHD Grant HD01994 to Haskins Laboratories. Correspondence should be addressed to D. J. Dekle, Department of Psychology, Dartmouth College, Hanover, NH 03755.

come because, as we have already pointed out, they concern the phonetic, not the lexical, properties of words as signaled acoustically and optically. For this reason, we were prompted to study Easton and Basala's findings more closely. Our examination did suggest a reason why Easton and Basala might have underestimated the role of visual information in spoken-word perception, and our experiments attempt to test our hypothesis.

In Easton and Basala's (1982) Experiment 1, standard lipreading (SLR) and dubbed lipreading (DLR) tests were performed with two groups of subjects. In the SLR, subjects identified mouthed words with the soundtrack turned off. In the DLR, they identified mouthed words with the soundtrack on. Subjects were tested twice, once with an identification task in which they wrote down the mouthed word and once with a multiple choice task in which they chose the mouthed word from among five alternatives. A third, control, group of subjects was instructed to watch the videotape with the sound on but to identify the word on the soundtrack. Five dubbing conditions were employed: (1) Same—In this condition, the mouthed word was the same as the word dubbed onto the soundtrack (e.g., *wild*[mouthed]-*wild*[dubbed]); (2) Initial—The mouthed word had the same initial visual information as the dubbed word (e.g., *face-fame*); (3) Final—The mouthed word had the same final visual information as the dubbed word (e.g., *teeth-mouth*); (4) Both—The mouthed word had the same initial and final visual information as the dubbed word (e.g., *word-whirl*); and (5) Neither—The mouthed word differed visually in every way from the dubbed word (e.g., *feel-roam*). There were two sets of word pairs in the initial and final groups. In one, the visual discrepancy was marked (as in *face-fame* and *teeth-mouth*). In the other, it was "obscure" (as in *buzz-bunch* and *chime-time*).[1]

A major finding of Easton and Basala's (1982) Experiment 1 was that discrepant visual speech information specifying complete words appeared to exert little or no effect on auditory speech identification. Conversely, discrepant acoustic information exerted a strong effect on lipreading. Subjects instructed to report the word they heard on the DLR were correct on 99% of trials. In contrast, subjects instructed to identify or select the visibly mouthed word from among alternatives were considerably less accurate with the sound on (DLR) than with it off (SLR). On the identification test, nearly a third of the lipread words were identified as the word on the soundtrack, while another, smaller, percentage of responses reflected integrations of optically and acoustically signaled phonetic information.

Our speculation that Easton and Basala (1982) may have underestimated an influence of visual information on judgments of heard words derives from our examination of their stimulus materials. They did not provide a list of all 30 pairs of words in their tests, but their Table 1 provides 9 of the 15 monosyllabic word pairs. (Another 15 pairs were spondaic disyllables.) Pair members in their *both* category, for example, the sample pair *word-whirl*, are almost certainly optically indistinguishable. Their vowels are the same, and their initial and final consonants have the same places of articulation. Although /d/ and /l/ differ in manner of articulation, manner and voicing distinctions are largely invisible (see Summerfield, 1987, for a review). Accordingly, the speaker's visible vocal tract movements for *word* are wholly consistent with *whirl*, the word on the soundtrack. As a result, when subjects were instructed to report the heard word, they would be expected to report *whirl* because the optical information was not discrepant with the auditory information. In the lipreading test, they also should have reported *whirl* because the soundtrack provided *whirl* as a candidate word that was consistent with the visible vocal-tract movements. The *both* condition, therefore, should not have (and did not) result in a McGurk effect but rather a strong reverse effect. The same outcome is likely on *initial* and *final* trials where the discrepancy between the mouthed and soundtrack words was obscure because, at least in the examples that Easton and Basala provided, the words had visibly the same places of articulation.

We cannot determine the proportion of trials on which Easton and Basala's (1982) words were not discrepant in this way. However, from their examples, we speculate that all of the *both* pairs and approximately half of (that is, the "obscure") *initial* and *final* pairs failed to provide discrepant place information. This leaves only the *neither* pairs and the remaining *initial* and *final* pairs that could have given rise to a McGurk effect. Those trials also failed to do so, perhaps for a different reason. Consider the sample *final* trial, *teeth-mouth*. For the initial consonant, subjects should have heard /n/, an integration of place information from the mouthed word and voicing and manner information from the soundtrack word (MacDonald & McGurk, 1978). The heard word should then be *nouth*, a nonword. In the identification test, subjects were instructed to identify "words," and the closest word to the predicted McGurk percept is *mouth*. In the multiple-choice test, *nouth* was not a response option. Of the nine sample trials that Easton and Basala offer, just one pair's expected McGurk outcome, when it is different from the soundtrack word, is a real word (*face-fame: feign*).[2]

Based on the foregoing analysis, our expectation is that with a different set of auditory-visual word pairs having consonants with phonetic properties shown to give rise to McGurk effects on nonwords, a McGurk effect should be obtained. With visible vocal-tract movements discrepant from those signaled by the acoustic information, perhaps the lipreading effects will be correspondingly diminished because the soundtrack will no longer offer a candidate word for lipreaders that is wholly consistent with what they see. Our experiments were designed to test these possibilities. (Refer to Table 1 for a schematic overview of Experiments 1-3.) Experiment 1 focused on the McGurk effect itself.

## EXPERIMENT 1

In Experiment 1, we tested for a McGurk effect with real words. Stimulus materials were selected so that optically and acoustically signaled place-of-articulation infor-

**Table 1**
**Schematic Overview of Experiments 1–3**

| Response Required | Task (Target Dimension) | |
|---|---|---|
| | Hearing (McGurk) | Lipreading (Reverse McGurk) |
| Multiple choice | TV on vs. TV off<br>Exp. 1 "McGurk effect" | Sound on vs. sound off<br>Exp. 2 "Reverse McGurk effect" |
| *Same–different* discrimination | TV on vs. TV off<br>Exp. 3 "McGurk effect" | Sound on vs. sound off<br>Exp. 3 "Reverse McGurk effect" |

mation was discrepant in an effort to maximize an effect of visual information on perception of the heard words.

## Method

**Subjects.** The subjects were 33 undergraduates at Dartmouth College who participated for course credit in an introductory psychology course. Eighteen subjects participated in a condition in which the TV monitor was on ("view"), and 15 participated in a condition with the TV monitor off ("no view"). All had normal hearing, had normal or corrected-to-normal vision, and were native speakers of English.

**Stimulus materials.** On the test videotape, there were five instances of each of nine trial types. The trial types are listed in Table 2. Of them, four had word-initial /b/ on the soundtrack paired with word-initial /v/ on the videotape, with /v/ the expected heard (McGurk) consonant (MacDonald & McGurk, 1978). The other five trial types had /m/ on the soundtrack paired with either /d/ or /g/ on the videotape, with /n/ the expected McGurk consonant.[3] All audio, video, and expected McGurk stimuli were real English words.

The model on the videotape (D.J.D.) was filmed saying multiple tokens of each video and soundtrack word. For purposes of dubbing, the first fluent and correct token of each video and audio word was selected. The dubbing was accomplished by filtering one token of each of the nine auditory words at 10 kHz, digitizing them at a sampling rate of 20 kHz, and storing them on a computer disk. Words starting with nasals were edited slightly to shorten the nasal murmur. This was done to improve their apparent synchrony with the stop consonants on the video display. To make the dubbed trials, the video signal from the original tape, which was played on one VCR, was recorded onto a new videotape on a second VCR while the audio signal from the original tape triggered a voice key interfaced with the computer. On receiving a signal from the audio track of the original videotape, the computer output a designated sound file to the second videotape, thereby dubbing it. Earlier estimates of the latency between voice-key triggering and sound-file output are 9 msec (Rosenblum & Fowler, 1991).

The new dubbed tape was edited to make a final tape consisting of 5 randomized instances of each of the 9 trial types in Table 2.

**Table 2**
**Soundtrack (Auditory) and Video (Visual) Word Pairs Used in Experiment 1 With Their Expected McGurk Integration; These Words Also Constituted the Three Response Options Offered to Subjects on Their Answer Sheets**

| Auditory | Visual | Expected McGurk Response |
|---|---|---|
| bat | vet | vat |
| bet | vat | vet |
| bent | vest | vent |
| boat | vow | vote |
| might | die | night |
| mail | deal | nail |
| mat | dead | gnat |
| moo | goo | new |
| met | gal | net |

Approximately 10 sec of black videotape separated each of the 45 trials.

**Procedure.** Two different between-group test conditions were devised: (1) a view condition with the TV monitor on and (2) a no-view condition with the monitor off. (The soundtrack was played over a speaker placed close to the TV.) In both conditions, the subject selected the word he/she heard on each trial from a set of three randomized alternatives corresponding to the auditory dub (e.g., *mat*), the visibly mouthed word (*dead*), and the expected integrated (McGurk) word (*gnat*). The response alternatives for each trial were listed on an answer sheet, and the subject was instructed to circle his/her choice, guessing if necessary.

The subjects sat 15 ft from a 21-in. TV screen and were tested in groups of 1–3. The soundtrack was played at a comfortable listening level in a quiet, but not soundproofed, room. In the view condition, the subject was told to watch the TV but to select the word he/she *heard*, not necessarily the one he/she saw mouthed, guessing if necessary. In the no-view condition, the subject was told to select the word he/she heard, guessing if necessary.

## Results and Discussion

In the no-view condition, the subjects were very accurate in reporting what they heard the speaker say. In that condition, 97% of the responses chosen were auditory selections, 3% were McGurk responses, and 0% were visual responses. In the view condition, 17% of responses were the auditory selections, 79% were the McGurk responses, and 4% were the visual responses. For purposes of analysis, the percentages of the responses falling into the two response categories of greatest interest and popularity (the auditory and McGurk integration responses) were subtracted to give a difference score. The score was large and positive in the no-view condition (94% on average) and large and negative in the McGurk condition (−62% on average). In one-tailed subjects-and-items $t$ tests (with $\alpha$ set to .017 on each of the three nonorthogonal tests we perform), the difference between these difference scores was highly significant [$t_1(31) = 17.72, p < .0001$; $t_2(8) = 43.25, p < .0001$]. Clearly, the presence or absence of information about the words on the video display had a strong effect on judgments of heard words.

Paired one-tailed $t$ tests of each difference score against zero showed that, as expected, there were significantly more auditory than McGurk responses in the no-view condition [$t_1(14) = 46.78, p < .001$; $t_2(8) = 21.89, p < .002$] and more McGurk than auditory responses in the view condition [$t_1(17) = 7.96, p < .001$; $t_2(8) = 19.52, p < .001$].

We ascribe the marked difference in our outcome as compared with that of Easton and Basala (1982) to our selection of stimulus materials. To obtain as strong a McGurk effect with words as with nonsense syllables,

place-of-articulation information for consonants must be discrepant across the modalities. In the next experiment, we look for a reverse McGurk effect, which would be an effect of the soundtrack word on subjects' judgments of the visible word.

## EXPERIMENT 2

It is likely that the same stimulus changes that resulted in a McGurk effect for the real words in Experiment 1 may also reduce the effect of the soundtrack on lipreading judgments (see our example of *word–whirl* above) as compared with the effect found by Easton and Basala (1982). The reduced effect is likely to occur because the soundtrack will no longer offer the subject a candidate word of the language that is wholly consistent with the visibly mouthed word. To determine whether there was a reverse McGurk effect with our stimuli, we used the videotape of Experiment 1 to obtain lipreading judgments.

### Method

**Subjects.** The subjects were 36 undergraduates from the same population as those who participated in Experiment 1.

**Stimulus materials.** The same apparatus, videotape, and answer sheets as those in Experiment 1 were used in this experiment.

**Procedure.** Two conditions were devised for Experiment 2. One group of 18 subjects viewed the same dubbed videotape from Experiment 1 with the TV picture and sound on, and another group viewed the same videotape with the TV picture on but the sound turned off. Both groups were asked to circle the word they *saw* the model say. We used the same answer sheet as in Experiment 1; accordingly, the subjects selected among audio, visual, and McGurk responses (see Table 2 above).

### Results and Discussion

In this experiment, in contrast to Experiment 1, the subjects were instructed to circle the word that best matched what they *saw* the speaker say. With the sound off, 69% of the responses corresponded to the video selection and 31% to the McGurk choice. With the sound on, the corresponding percentages were 55% and 38%, respectively; on the remaining 7% of the trials, the subjects selected the auditory response option.

As in Experiment 1, we converted the percentages of responses falling into the two main response categories—here, visual and McGurk—to a difference score by subtracting them. This gave a positive difference score in the no-sound condition (38% on average) and, unexpectedly, in the sound (17%) condition as well, signifying that numerically more visual than McGurk responses were selected in both conditions. A one-tailed $t$ test ($\alpha = .017$) of the difference between conditions was significant with subjects as the random variable and marginal with items as the random variable [$t_1(34) = 2.23, p = .015$; $t_2(8) = 2.05, p = .035$]. Paired $t$ tests of the difference scores against zero revealed significantly more visual than McGurk response selections in the no-sound condition [$t_1(17) = 10.03, p < .001$; $t_2(8) = 2.54, p = .017$].[4] However, the (unpredicted) tendency for visual responses to dominate in the sound condition was weak and nonsignificant on two-tailed tests [$t_1(17) = 2.12, p = .047$; $t_2(8) = 1.21, p = .23$].

Overall, there was at most a weak reverse McGurk effect. Its magnitude is difficult to compare with that of Easton and Basala (1982) because the baseline levels of performance in their multiple-choice tests were different from ours. However, it is weaker in its outcome as evaluated statistically, and it is considerably weaker than the cross-modal effect in the present Experiment 1. In analyses of variance comparing the effect of condition across the two experiments, there was a highly significant interaction with experiment in both the subject and item analyses [$F_1(1,65) = 194.41, p < .001$; $F_2(1,8) = 141.6, p < .001$], reflecting the considerably stronger influence of lipread information on judgments of heard words than of the reverse.

Our conclusions, so far, are the following. Under conditions in which the literature suggests that McGurk effects occur on nonwords, they occur on words as well; under those same conditions, influences on lipreading (reverse McGurk) are weaker than McGurk effects. It does not follow, of course, that this latter finding can be generalized to speech perception outside the laboratory. Since the conditions that allow us to observe bidirectional cross-modal influences do not occur there, we cannot draw inferences from our own findings or those of Easton and Basala (1982) regarding any general differences in the relative magnitudes of the two cross-modal influences outside the lab. Looking across Easton and Basala's findings and our own, we can conclude that, in the laboratory, the effect that looks larger when the modalities are placed in conflict depends on the phonetic, not the lexical, properties of the stimulus items in each modality. Finally, an inference that we do wish to draw concerning normal conditions of speech perception is that, given the opportunity, listeners do acquire phonetic information optically. This, as noted in the introduction, is expected in the motor theory of speech perception, the direct-realist theory, and Massaro's (1987) fuzzy-logical model.

## EXPERIMENT 3

Effects of acoustically specified words on lipread words were marginal in Experiment 2 in contrast to the strong reverse influence obtained in Experiment 1. There was a further difference, more difficult to quantify, in the two experimental tests. The McGurk effect itself gives rise to a marked change in the perceiver's phenomenal experience of hearing a word. An acoustically signaled word sounds different paired with an appropriately chosen, phonetically discrepant video display than it sounds unpaired. Our own phenomenal experiences of seeing mouthed words were at most subtly affected by the presence or absence of a soundtrack. It seemed possible to us (on the basis of only these phenomenal experiences, not any theoretical reasons as to why the effects should differ) that the strong McGurk effect and the weak lipreading effect might have different origins. The McGurk effect may be a true perceptual effect, whereas the other

reflects a bias to guess in the direction of the heard word because the optical information for a word is ambiguous. If such a bias were, in fact, the source of the weak lipreading effects we found in Experiment 2, we thought it possible that they would disappear were subjects given an easier task than one of identification of words in isolation. Accordingly, in the present experiment, we asked subjects to compare a sequential pair of mouthed words and to decide whether they were the same or different words (lipreading condition). Subjects participated in two versions of the *same–different* task, one in which the soundtrack was turned on (lipreading bimodal) and one in which it was off (lipreading unimodal). On some trials when the mouthed words were the same, the dubbed words on the soundtrack were selected so as to promote a McGurk experience of hearing the same word repeated; on other trials, the expected heard words were different. This allowed us to determine whether the ability to detect that the mouthed words were the same would change depending on the words subjects experienced hearing. Other subjects took analogous *same–different* tests (hearing condition) and judged whether two words they heard were the same or different, with the TV monitor either on (hearing bimodal) or off (hearing unimodal).

## Method

**Subjects**. The subjects were 24 undergraduates from the same population as those who participated in the previous experiments. Twelve subjects participated in the hearing (McGurk) condition, and the other 12 subjects participated in the lipreading (reverse McGurk) condition.

**Stimulus materials**. In the bimodal tests, there were five sets of three dubbed word pairs in the lipreading condition and another five sets in the McGurk condition. (The same materials were used in the unimodal conditions except, of course, that the soundtrack words were absent in the lipreading test and the video display was absent in the hearing test.) Within a set, the triads included one item meant to evoke a "different" judgment, one meant to evoke a "same" judgment, and one meant to evoke a "same" judgment only if the subjects were able, as instructed, to ignore information in the other sensory modality. In the lipreading test, in which subjects were asked to judge whether mouthed (visible) words in a pair were the same or different, words of a pair on the soundtrack were always different. In one trial type of the three, a word pair such as "bent bun" both on the soundtrack and on the video display should provoke a "different" response, particularly if the subjects are not able to ignore information on the soundtrack. The video word pair "bent bent" should provoke a "same" response if it is paired on the soundtrack with words that do not change the perceived identity of the mouthed words. We selected acoustic "bent vent" for this purpose, needing to have different words on the soundtrack so that, by themselves, the soundtrack words would not bias a "same" judgment and speculating that the voiced frication for the /v/ in the context of visible bilabial closure would be heard as prevoicing for /b/. (We checked this and other guesses using an identification test to be described below.) In the lipreading condition, the third word pair of a triad, like the second, had visibly the same word spoken in succession—in our example, "bent bent." In this case, however, the soundtrack word pair was selected so as to change the perceived identity of the spoken words—here, the soundtrack presented "bent bum." The McGurk integration of visible "bent bent" and acoustic "bent bum" should be "bent bun," with the place of articulation of the final consonant of the second word being supplied by the video display but the vowel and consonant manner supplied by the soundtrack.

To summarize, in each triad, there were two critical trial types in which the same mouthed words were presented in succession. In one of them, the heard words were different but were selected so as to ensure that the McGurk integration would be the same as the mouthed sequence in one case and to ensure that it would be different in the other case. Does the perception that a speaker said "bent bun" make mouthed "bent bent" look different, and different from the way it looks when the perception is that the speaker said "bent bent"?

An analogous set of five triads was devised for a hearing test. In this test, the subjects were to judge whether the words in a pair on the soundtrack were the same or different. Again, one item of a triad presented different words both acoustically and optically and were expected to provoke "different" judgments. An example is "bent vent" presented both optically and acoustically. In a second member of each triad, the words on the soundtrack were the same, for example, "bent bent." In this case, the synchronized words on the videotape were different so as not to bias a "same" response themselves, but they were selected so that they would not shift the perceived heard words away from the words on the soundtrack. In the example, the visible words were "bent bat" and the expected heard sequence should be "bent bent." In the third item of each triad, the same soundtrack words (again "bent bent" in the example) were paired with words on the video display that should have changed the words as heard. In the example, mouthed "bent vent" should lead subjects to identify the spoken words "bent bent" as "bent vent."

As on the lipreading test, the major question of interest here was whether the same words in the judged modality would be classified differently depending on what they were paired with in the other modality. In this case, on the basis of our own past experience and that reported in the literature, we could be confident that the answer would be yes.

The model for the videotape of Experiments 1 and 2 also served as a model for the *same–different* test. She was recorded saying pairs of words in close succession (average stimulus onset asynchrony [SOA], approximately 590 msec in the lipreading test [reverse McGurk] and 690 msec in the hearing test [McGurk]; average word duration, 448 msec in the lipreading test [reverse McGurk] and 428 msec in the hearing test [McGurk]). Stimuli were dubbed in the same way as described in Experiment 1, except that two words had to be dubbed on each trial. Each word of a *same–different* pair on the original videotape soundtrack separately triggered dubbing of a new soundtrack word by the computer, as described for Experiment 1. However, in this case, a different voice (Fowler, recorded in a sound-attenuating room) was dubbed onto the videotape because the original sound recording was noisy. Multiple tokens of each spoken word were recorded, from which two were selected for recording. On trials on which the same word was presented on either the soundtrack or the video display, the word tokens were different productions of the same word.

In the final lipreading and hearing-test tapes, each of the 15 trial types for a given test (that is, 5 sets of 3 audio-visual word pairs) appeared twice, randomized, once in an AB order and once in a BA order, with 5 sec between trials. One trial from the lipreading test was inadvertently left out, leaving just 29 trials on that test tape.

**Procedure**. Two conditions were devised, a McGurk condition and a reverse McGurk condition. Two groups of 12 subjects each participated in three lipreading tests or three hearing tests. Within each group, the order of the tests (identification, unimodal, bimodal) was counterbalanced, with 2 subjects experiencing each of the six possible orderings of the three tests. In devising our stimulus triads, as already noted, we had to guess at the perceptual result of many of the audio-visual pairings we used. Therefore, all subjects in both the lipreading and the hearing-discrimination conditions participated

in an identification test in which they watched the videotape for their condition with the sound on and attempted to identify the spoken word pairs they heard on each trial. We told the subjects that sometimes the words might sound strange and that they should attempt to spell out exactly what they heard on each trial even if it was a nonword. We used the results of the identification test for each subject to eliminate trials on the *same–different* tests for which their perceptual reports were different than we had expected.

In the lipreading test (reverse McGurk), the remaining two tests were *same–different* discriminations in which the subjects were to indicate whether the words of a pair they saw the speaker say were the same or different words. The subjects took the test once with the sound on and once with it off.

A different group of subjects took the identification test and two analogous *same–different* tests on the hearing-test tape (McGurk). The instructions on the identification test were the same as those given to lipreading (reverse McGurk) subjects. The instructions to the subjects on the *same–different* task were to make their classifications on the basis of what they heard, not what they saw when the TV was on. They took the test once with the TV on and once with it off.

## Results and Discussion

The subjects' responses on the identification test were used to eliminate trials on which their perceptual reports were discrepant with our expectations. This was important in order for us to compare the two trial types on each test in which the words of a pair in the judged modality were the same (in the examples used above, "bent bent" in the video display of the lipreading test and on the soundtrack of the hearing test) but the expected integrated percepts were either of the same words (video "bent bent" paired with soundtrack "bent vent"; soundtrack "bent bent" paired with video "bent bat") or of different words (video "bent bent" paired with soundtrack "bent bum"; soundtrack "bent bent" paired with video "bent vent"). The comparisons are only valid if the subjects heard the same words and different words in the *same* and *different* conditions, respectively. Accordingly, we eliminated from the analysis both of the bimodal (sound condition of the lipreading test and view condition of the hearing test) and the unimodal (sound off or TV off) *same–different* tests, trials on which, on the identification test, a subject had either reported the same word repeated on a *different* trial or different words on a *same* trial. We also eliminated trials on which a subject reported consonant sequences (such as *mnet* for video *met* compared with audio *net*). On average, 23% of trials were eliminated in the lipreading test and 17% in the hearing test.

An index of the subjects' ability to do the *same–different* task is provided by their performance on *different* trials where all of the evidence subjects have—on bimodal trials, the evidence includes the mouthed words, the acoustic words, and the McGurk-experienced words—should lead them to respond "different." In the lipreading test, performance was poor. In the unimodal test, 54% of responses were correct (that is, with the sound off). In contrast, with the sound on, performance rose to 77% [$t_1(11) = 3.39, p = .006$; $t_2(4) = 1.82, p = .14$]. In the hearing test, performance was almost perfect, with 99% correct responses on the unimodal test and 100% correct on the bimodal test.

On the unimodal *same–different* tests—on the lipreading test with the sound off or the hearing test with the TV off—the two categories of "same" responses collapse. That is, the difference between them depends on the differences expected to be caused by information in the other modality, present only in the bimodal test. On unimodal *same* trials of the lipreading test, performance averaged 68% correct. It averaged 99% on the hearing test.

The questions of central interest concerned the *same* trials of the bimodal tests. Will the subjects' tendency to judge a *same* pair of words as visibly the same word (lipreading test) or as audibly the same word (hearing test) be affected by any integration of information from the to-be-ignored modality? That is, will the subjects be less likely to judge optical "bent bent" as the "same" if they judge the speaker to have said "bent bun" (optical "bent bent" paired with soundtrack "bent bum") than if they judge the speaker to have said "bent bent" (optical "bent bent" paired with soundtrack "bent vent")? Analogously, will the subjects be less likely to judge soundtrack "bent bent" as the "same" if they identify the words as "bent vent" (soundtrack "bent bent" paired with optical "bent vent") than if they identify them as "bent bent" (soundtrack "bent bent" paired with optical "bent bat")?

The answer to these questions is "yes." On the lipreading (reverse McGurk) test, the subjects judged *same* sequences as "same" on 36% of trials on which they had identified the spoken words differently and on 81% of trials on which they had judged them as the same word repeated. The difference between the conditions is significant [$t_1(11) = 5.90, p < .0001$; $t_2(4) = 5.0, p = .0075$]. On the hearing (McGurk) test, the corresponding percentages are 25% and 100% [$t_1(11) = 8.53, p < .0001$; $t_2(4) = 15.22, p < .0001$].

For two reasons, the results on the hearing and lipreading tests cannot be assumed to be comparable. First, lipread information for words is simply poorer than is acoustic information (as indexed, for example, by the fact that deafness provides a considerably more severe barrier to spoken-language learning than does blindness). Accordingly, the lipreading test will be harder than the hearing test. As one reviewer pointed out, intuitively, that difference would work in favor of finding a *stronger* influence of the soundtrack on lipreading judgments than the reverse effect. That is, good acoustic information for a word should be little challenged by poor optical information, whereas ambiguous lipread information might be strongly influenced by a clearly identified acoustic signal. Second, however, the stimulus materials in the two conditions necessarily were different because, to a first approximation (Summerfield, 1987), the phonetic information that is influential optically is complementary to that which is influential acoustically. That the materials are different is not in itself bad. The materials in Experiments 1 and 2 were the same but were probably biased in favor of a strong McGurk finding; those in Easton and Basala (1982) were the same in lipreading and McGurk conditions but were biased toward a strong lipreading effect. The diffi-

culty is that, whereas there is a literature on the effective conditions for producing a McGurk effect, there is almost none that explores conditions for producing strong and weak influences of a dubbed soundtrack on lipreading judgments. We had to guess, and our guesses cannot ensure that the tests in the two modalities are comparable. Hence we will not compare them statistically; we included the hearing test only to ensure that the McGurk effect remains in the *same–different* testing conditions we used.

Overall, the results of the experiment did not further differentiate the lipreading and hearing effects as we had speculated it might. With our stimuli, the McGurk effect is qualitatively similar to, but still numerically stronger than, the effect of sound on lipreading.

## GENERAL DISCUSSION

Our three experiments have explored some conditions under which cross-modal influences occur in reports of perceived words. In contrast to the findings of Easton and Basala (1982), we found a strong influence of lipread words on judgments of heard words on a synchronous soundtrack. The critical difference in outcome, we believe, lies in our selection of stimulus materials. To obtain a McGurk effect for real or nonsense words, it is necessary for the visually and acoustically signaled words to be discrepant on a phonetic dimension that perceivers can detect. Earlier research (e.g., MacDonald & McGurk, 1978) had shown that discrepancies in certain places of articulation of consonants foster the most compelling cross-modal integrations. Easton and Basala's stimuli appear largely not to provide visible place discrepancies, and so, expectedly, McGurk integrations are absent.

Our findings support those of Easton and Basala (1982) in obtaining a reverse cross-modal influence of a heard word on judgments of lipread words. Our effect may have been smaller than theirs, and in contrast to theirs, it was also weaker than the McGurk influence of lipread words on judgments of heard words. We ascribe this difference also to differences in selection of stimulus materials. In Easton and Basala's stimulus materials, in many cases, the lipread and heard words were not noticeably discrepant. This not only eliminates the possibility of finding a McGurk integration on those pairs but also fosters the appearance of a very strong reverse cross-modal influence because the word on the soundtrack provides a candidate word that exactly fits the lipread vocal-tract gestures.

With cross-modal consonantal places of articulation detectably discrepant, as in our stimulus materials, there were two consequences: (1) there was now the possibility of McGurk-type blends in judgments of the heard word, and (2) in the other direction, the soundtrack no longer provided a candidate word that was wholly consistent with the video word. This latter consequence permitted an influence of the soundtrack on lipreading judgments more analogous to the McGurk influence of lipreading judgments of heard words. We found such an influence, although it was weak.

Our third experiment was an attempt to provide a more sensitive examination of the influence of sound on lipreading judgments. We wanted to know, analogous to the McGurk effect, which influences the phenomenal experience of hearing a word, whether the reverse influence affects what perceivers experience seeing. We used a *same–different* discrimination test with a short SOA between words in an effort to determine whether subjects would judge the same lipread sequence differently depending on whether the words on the soundtrack were the same or different. Subjects did judge them differently, albeit less strikingly than on the analogous hearing test. Our tentative conclusion from this experiment is that the cross-modal influence of heard words on lipread words is of the same general sort as the McGurk influence of lipread words on spoken ones.

Returning to the McGurk effect itself, we conclude from our findings in comparison with those of Easton and Basala (1982) that the conditions under which the effect occurs or fails to occur must be described phonetically, not lexically. Visual information for phonetic contrasts is influential for those phonetic properties that are observable visually. Information for some places of articulation are particularly visibly apparent; complementarily, acoustic information for place is generally quite labile (Summerfield, 1987). When the two information sources conflict, the more reliable source dominates. As for speech perception outside the laboratory, the McGurk findings suggest that listeners do use optical sources of information when it is available. If they did not, their ability to use it under laboratory conditions would be surprising.

As noted in the introduction, the finding that a McGurk effect can occur on words as well as nonwords is comforting to the set of speech-perception theorists who have attempted to explain the effect. According to motor theorists (e.g., Liberman & Mattingly, 1985) and direct realists (e.g., Fowler, 1986), listeners to speech hear the linguistically significant (phonetic) actions of a speaker's vocal tract that give rise to acoustic speech signals. To the extent that information about those actions is conveyed optically, it will be used perceptually as well. There is nothing about the lexical status of a sequence of phonetic segments that will alter the tendency to use optical information. (The motor theory and direct-realist theory disagree about how the gestural information is perceived—via analysis by synthesis or directly—and about whether gestural recovery in speech perception makes speech perception "special" or wholly unspecial [e.g., Fowler, 1991].) In Massaro's (1987) fuzzy-logical theory, syllables are perceived by matching stimulus information to prototypes in memory. Associated with syllable prototypes are features that can signal the syllable when it is spoken. Features are optical as well as acoustic. Again, there is nothing about the lexical status of a syllable that would change

the effectiveness of optical speech input. In short, none of the theories is challenged by the conditions under which the McGurk effect has been found to occur to date.

## REFERENCES


CAMPBELL, R., GARWOOD, J., FRANKLIN, S., HOWARD, D., LANDIS, T., & REGARD, M. (1990). Neuropsychological studies of auditory-visual fusion illusions: Four case studies and their implications. *Neuropsychologia*, **28**, 787-802.

EASTON, R. D., & BASALA, M. (1982). Perceptual dominance during lipreading. *Perception & Psychophysics*, **32**, 562-570.

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

FOWLER, C. A. (1991). Auditory perception is not special: We see the world, we feel the world, we hear the world. *Journal of the Acoustical Society of America*, **89**, 2910-2915.

FOWLER, C. A., & DEKLE, D. J. (1991). Listening with eye and hand: Cross-modal contributions to speech perception. *Journal of Experimental Psychology: Human Perception & Performance*, **17**, 816-828.

LIBERMAN, A. M., & MATTINGLY, I. G. (1985). The motor theory revised. *Cognition*, **21**, 1-36.

MACDONALD, J., & MCGURK, H. (1978). Visual influences on speech perception processes. *Perception & Psychophysics*, **24**, 253-257.

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

MCGURK, H., & MACDONALD, J. (1976). Hearing lips and seeing voices: A new illusion. *Nature*, **264**, 746-748.

ROSENBLUM, L. (1989). *Effort perception of speech and nonspeech events: An audio-visual investigation*. Unpublished doctoral dissertation, University of Connecticut, Storrs, CT.

ROSENBLUM, L., & FOWLER, C. A. (1991). An audio-visual investigation of the loudness/effort effect for speech and nonspeech perception. *Journal of Experimental Psychology: Human Perception & Performance*, **17**, 976-985.

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

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


## NOTES

1. In word pairs for which the discrepancy is "obscure," the visible difference between the target consonants was subtle at most. In the examples in the text, the consonantal constriction location for /č/ (the alveopalatal consonant spelled "ch") is slightly behind that of /t/ and /z/ (alveolar). However, both places of articulation reflect constrictions by the tongue blade with the hard palate just behind the teeth; accordingly, the small difference in place of articulation is not likely to be visible as such. (The /n/ in "bunch" will be produced with /č/'s place of articulation in that context.) In addition, /z/ is voiced, whereas /č/ is unvoiced, but that is an invisible difference in the state of the vocal folds in the larynx. As for manner of articulation, /č/ is an affricate—a stop constriction released as a fricative—whereas /t/ is a stop and /z/ is a fricative. These manners of articulation differ very subtly in the ways in which the tongue achieves (affricate vs. fricative) or releases (affricate vs. stop) the consonantal constriction; therefore, these differences are also likely to be invisible. The phoneme /č/ may be distinguished from /t/ by a degree of lip rounding (Summerfield, 1987), but rounding may also occur for /z/.

2. In a review of our paper, Easton (personal communication, August, 1991) suggested that our arguments fail to apply to the spondaic visual-auditory word pairs used by Easton and Basala (1982). Examples of these pairs were not provided in the paper, but Easton provided 5 of the 15 pairs in his review. They are "notebook-cufflink," "northwest-cowgirl," "keepsake-washroom," "pastime-trespass," and "oatmeal-doorbell." He commented that the medial consonants particularly did differ in place of articulation. Yet no McGurk effects were found for these items, either, "presumably because of their extended temporal structure and resultant auditory information." However, we believe that the reason why these items failed to show a McGurk effect is the same as the reason that we speculated why "teeth-mouth" failed to show an effect. Consider the visual-auditory pair "notebook-cufflink." Given appropriate temporal alignment of the words, the initial consonant should be heard as /t/ (cf. MacDonald and McGurk, 1978)—that is, as an alveolar (from the videotape) voiceless stop (from the soundtrack). Analogously, the second consonant should be heard as /s/, the third as a bilabial voiced continuant, perhaps /w/, and the final consonant cluster should be heard as such—that is, as /ŋk/. Perception of heard vowels generally is little affected by dubbing (but see Summerfield & McGrath, 1984, for evidence of some integration); accordingly, the vowels should be those of "cufflink," giving something like "tusswink" as the perceived sequence. This was not a response option in the multiple-choice test used by Easton and Basala, and the subjects were instructed to report words in the free-identification test. The subjects, then, did not have an opportunity to reveal any McGurk integration they may have experienced. The other word pairs above are subject to the same account—with the possible exception of "oatmeal-doorbell" where the McGurk integration may be "doorbell" itself. (In that pair, the only visible "place discrepancy," if there is any at all, may occur at word onset, where there is no initial consonant in "oatmeal.")

3. One reviewer questioned the generality of our findings given the small number of auditory-visual pairings we used. Research on the McGurk effect that has used nonwords has, to a degree, explored the generality of the findings across phonetic contexts. Our aim here is not to replicate or to extend that research in the phonetic realm but rather to show that the kinds of visual-auditory pairings that give rise to a strong McGurk effect with nonwords also give rise to a strong McGurk effect with words, presumably for the same reason. In any case, a recent study by Campbell et al. (1990), unknown to us when we designed our study and designed for a different purpose, has, without commenting on Easton and Basala's (1982) claims, used words involving other pairings than our own and has obtained a strong McGurk effect with them.

4. A reviewer suggested that the visual and McGurk response options may be word pairs that, when lipread, are not noticeably discrepant; however, they are noticeably discrepant, as the results of this test indicate.

(Manuscript received June 25, 1991;
revision accepted for publication November 4, 1991.)