Perception & Psychophysics
1993, 54 (3), 406-416

Visual influences on auditory pluck and bow judgments

HELENA M. SALDAÑA and LAWRENCE D. ROSENBLUM
University of California, Riverside, California

In the McGurk effect, visual information specifying a speaker's articulatory movements can influence auditory judgments of speech. In the present study, we attempted to find an analogue of the McGurk effect by using nonspeech stimuli—the discrepant audiovisual tokens of plucks and bows on a cello. The results of an initial experiment revealed that subjects' auditory judgments were influenced significantly by the visual pluck and bow stimuli. However, a second experiment in which speech syllables were used demonstrated that the visual influence on consonants was significantly greater than the visual influence observed for pluck-bow stimuli. This result could be interpreted to suggest that the nonspeech visual influence was not a true McGurk effect. In a third experiment, visual stimuli consisting of the words *pluck* and *bow* were found to have no influence over auditory pluck and bow judgments. This result could suggest that the nonspeech effects found in Experiment 1 were based on the audio and visual information's having an ostensive lawful relation to the specified event. These results are discussed in terms of motor-theory, ecological, and FLMP approaches to speech perception.

It has been demonstrated that visual information can influence auditory speech judgments. For instance, when an acoustic syllable is paired with a video tape of a speaker producing a different syllable, listeners sometimes report hearing what they actually see, or report hearing a blend of the visual and auditory signal (MacDonald & McGurk, 1978; McGurk & MacDonald, 1976). This "McGurk effect" is robust. It occurs even when subjects notice an incompatibility between the auditory and visual components of the syllables (Repp, Manuel, Liberman, & Studdert-Kennedy, 1983), and it occurs when subjects are told to base their judgments only on what is "heard" (Summerfield & McGrath, 1984). What is yet to be determined is whether an analogous effect might occur with nonspeech sounds.

Our research explores the issue of whether visual influences on identification judgments can occur with nonspeech events. Here, three theories of speech perception that include explicit accounts of audiovisual speech will be presented. This will be followed by a review of the relevant literature on sensory dominance. Finally, three experiments will be reported; they were designed to test for a nonspeech McGurk effect and to determine how this effect might compare to a speech McGurk effect.

Three Theories of Audiovisual Speech Perception

Three theories of speech perception have been used to account for McGurk findings. These include the motor theory of speech perception (Liberman & Mattingly, 1985), the direct-realist approach (Fowler, 1986; Fowler & Rosenblum, 1991), and the fuzzy logical model of perception, or FLMP (Massaro, 1987).

According to the motor theory (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; Liberman & Mattingly, 1985), speech perception is achieved by a speech module that takes in relevant speech information (e.g., acoustic, optic) and forms a hypothesis about the articulatory sequence of consonants and vowels that give rise to the acoustic signal. It is proposed that listeners are able to use an internal analogue of the speech-motor system to test this hypothesis (Liberman & Mattingly, 1985). This leads to an object of perception that is not the acoustic signal, but rather the intended phonetic gesture of the speaker. Regarding the McGurk effect, the motor theorists propose that subjects report hearing a coherent syllable (from a discrepant audiovisual presentation) because the perceptual object is neither the acoustic signal nor the reflected light but rather the intended articulatory event (Liberman & Mattingly, 1985). This gestural interpretation might also explain why observers often cannot distinguish the auditory and visual contributions of their percepts.

Given that the gestural explanation is specific to speech, the motor theorists would likely propose that a true McGurk effect would be special to speech. (A true McGurk effect for the motor theorist would seem to involve an unambiguous auditory token of one type, visually influenced to be "heard" as an unambiguous token of another, while the auditory and visual contributions of the

We gratefully acknowledge the assistance of Suzanne Cavanagh and Dave Williams, as well as the comments of Carol Fowler, Dominic Massaro, and two anonymous reviewers. This research was supported by an intramural grant from the University of California. Requests for reprints should be sent to L. D. Rosenblum, Department of Psychology, University of California, Riverside, CA 92521.

percept cannot be parsed; see Liberman & Mattingly, 1985.) However, the motor theorists would likely allow for (extramodular) visual influences of nonspeech stimuli that do not portray the striking power of the McGurk effect (i.e., with more subtle changes in heard tokens). These intermodal influences could occur at some higher cognitive processing stage where dimensions of nonmodular stimuli are thought to be associated with particular classes of events (Mattingly & Liberman, 1990).

The direct-realist theory is similar to the motor theory, in that it also proposes that the object of speech perception is the phonetic gesture of the speaker (Fowler & Rosenblum, 1991). The critical difference between the approaches lies in how the object of perception is thought to be recovered. According to the direct-realist theory, perceivers detect invariant information in media (light, air), which is lawfully structured by the objects and events in the environment, and may therefore specify those objects and events (Gibson, 1979; Michaels & Carello, 1981). As applied to speech perception, this theory proposes that there exists invariant information in the acoustic (optic, haptic) signal that specifies the vocal tract actions of the speaker. In the case of the McGurk phenomenon, there exists higher order optic and acoustic information that specifies a coherent distal event different from that specified by the auditory information alone. It is important to note that speech perception from the direct-realist perspective does not require a specialized speech module: speech is not perceived in a special way. Instead, all perceiving involves the perception of the distal event. Therefore, it is proposed by the direct realist that a true McGurk-type effect can occur—in principle—for both speech and nonspeech events. The degree of the influence, however, could be dependent on such dimensions as stimulus saliency and the general nature of the specified events.

The FLMP (Massaro, 1987) does not maintain that the perceptual objects are distal in nature. Instead, the FLMP explains McGurk effects by proposing that auditory and visual sources of information are evaluated independently and then integrated. The integrated representation is then matched against prototype descriptions in memory. According to the FLMP, these prototype descriptions are built up from a perceiver's previous experience of utterances, during which optical and acoustic cues have been detected. The listener selects the memory representation that has the most in common with the integrated syllable, and this is the syllable that the subject hears. This process of information integration is considered to be general and can apply to other types of stimuli beyond speech (Massaro, 1987). Accordingly, the FLMP proposes that, in principle, true McGurk-type effects could occur with nonspeech stimuli. The degree of this effect, however, would depend on the salience of the visual cues for the particular event (and/or the ambiguity of the auditory cues).

To summarize, although all three theoretical approaches might expect some visual influence on nonspeech tokens, only the direct and FLMP approaches would allow for influences as striking as those found for speech in the McGurk effect.

Examples From Sensory Dominance Research

There is evidence that visual information can influence perception of nonspeech stimuli. Rosenblum and Fowler (1991) used the McGurk procedure to demonstrate a similar patterning of visual influence on loudness judgments of speech syllables and hand-clapping. This finding does suggest that visual information can influence judgments of heard nonspeech. However, this effect differs from the standard McGurk effect in that it does not demonstrate a visual influence on event *identification*. In the standard McGurk effect, subjects report hearing a different event from that which is specified in the acoustic signal (e.g., a /ba/ becomes a /va/). In the loudness experiment, subjects reported hearing the same event (hand-clapping) change subtly along a single dimension (loudness).

A number of other findings involving nonspeech sensory integration have been reported in the literature. Many such studies have dealt with spatial location judgments. For example, it has been shown that auditory localization of a token can be influenced by a discrepant visual token (e.g., Bermant & Welch, 1976; Choe, Welch, Guilford, & Juola, 1975). Other studies have demonstrated a visual effect on haptic perception with properties such as size (Kinney & Luria, 1970), depth (Singer & Day, 1969), and curvature (Easton & Moran, 1978). Finally, recent studies have demonstrated an auditory influence on visual stimuli in perceived duration (Walker & Scott, 1981), temporal rate (Welch, DuttonHurt, & Warren, 1986), and number of visual events (O'Leary & Rhodes, 1984). Still, no studies have demonstrated a visual influence on judgments of nonspeech object or event identification that would seem more analogous to the McGurk effect.

In the following experiments, we attempted to test a nonspeech McGurk effect that involved event identification. The stimuli for the first experiment involved audio and video presentations of plucks and bows on a cello. Pluck and bow events were chosen for a number of reasons. First, previous research has shown that subjects are able to accurately identify and distinguish pluck- and bow-type sounds (e.g., Cutting & Rosner, 1974). Second, pluck and bow audio stimuli can be modified to produce an auditory continuum analogous to others that have proven useful for demonstrations of McGurk effects (e.g., Fowler & Dekle, 1991; Massaro & Cohen, 1983; Rosenblum & Fowler, 1991; Summerfield, 1979). Finally, the visual difference between pluck and bow events is obvious to any observer, so these events are good candidates for observing a visual influence.

EXPERIMENT 1

For the first experiment, visual plucks and bows were paired with tokens from a pluck-bow auditory continuum.

Method

Subjects. The subjects were 13 undergraduate students at the University of California, Riverside. All reported normal hearing and normal or corrected vision. As in previous studies involving pluck and bow stimuli, the subjects were not selected for their musical abilities (e.g., Cutting & Rosner, 1974).

Stimuli. A Panasonic PVS350 camcorder and a Shure SM57 microphone were used to record the initial stimulus tape. The actor was seated 5 ft in front of the camera. The camera focus was centered on the body of the cello. The recorded image showed the cello body from top to bottom. Since the actor was seated behind the cello, only his shoulders and right arm were clearly visible. Each visual event started with the actor's right arm out of the picture. The right forearm and hand then entered the picture either with or without the cello bow. For the bowed visual stimulus, the actor placed the bow on one cello string and played a single note (the note G; ~98 Hz). For the plucked visual stimulus, the actor placed one finger on the same string and plucked the same note. Both the pluck and the bow visual events lasted 3 sec.

The five-point auditory continuum was generated on a Compaq 386/25 computer. A good exemplar of a bowed token was digitally sampled (at 20 kHz) from the original video tape into a file on the computer. This original token was about 550 msec in duration. With a speech analysis software package (CSRE), a continuum was generated by cutting off the rising portion (time between onset and greatest stimulus intensity) of the sampled bow token in equal 20-msec increments (care was taken to edit the signal at zero crossings). (Although this procedure is different from the stimulus generation used by Cutting & Rosner, 1974, and others, such as Rosen & Howell, 1981, informal pilot experiments determined that this editing produced a convincing continuum of pluck- and bow-type sounds.) The endpoint *bow* stimulus had a rising portion of 100 msec and the endpoint *pluck* stimulus had a rising portion of 20 msec. Further editing was performed on the end of each token so that all five tokens were 450 msec in duration. For example, for the endpoint bow token with a 100-msec rising portion, 100 msec were cut off the end of the stimulus to attain a 450-msec token. For the endpoint pluck token with a 20-msec rising portion, only 20 msec were cut off the end of the stimulus to produce a 450-msec token. Consequently, the wave form at the end of each token was slightly different from the others. This difference in offset was not noticeable to the experimenters.

The edited audio files were then dubbed synchronously with the video pluck and bow presentations. This was accomplished by using a video player, a video recorder, and the computer interfaced with a sound-activated circuit. To dub each token, the original tape was played so that its video signal was output to the video recorder and its audio signal was output to the sound-activated circuit. Upon sensing an audio token, the sound-activated circuit signaled the computer to output an edited audio token to the video recorder. Thus, the video token of the original tape and an edited audio token were recorded simultaneously onto a second tape, resulting in a new synchronous audiovisual token. The lag time for dubbing was found to be no greater than 9.4 msec, well below the 80-msec range required for observers to detect an audiovisual asynchrony (McGrath & Summerfield, 1985).

Through the use of this procedure, each of the five audio stimuli was paired with each of the two video stimuli. The resulting 10 audiovisual tokens were then recorded onto a third presentation tape. Each of the dubbed tokens appeared on the presentation tape six times. The tokens were arranged in three blocks of 20 tokens each, so that each token appeared in a block two times. Each block was presented to subjects twice, producing a total of six block presentations. Four of the block presentations were audiovisual, and two were audio alone (with the video portion of the stimulus turned off). Therefore, subjects judged a total of 80 audiovisual presentations (10 audiovisual tokens × 8 times each) and 40 audio-alone presentations (5 audio-alone tokens × 8 times each).

Procedure. The presentation tape was shown to the subjects (2 or 3 at a time) in a quiet room at the University of California, Riverside. The subjects sat about 5 ft in front of a 19-in. television monitor. The audio channel of the tape was output through a small speaker that was located directly underneath the video monitor.

Pluck and bow identifications were in the form of graphic ratings (see, e.g., Rosenblum, 1989; Rosenblum & Fowler, 1991). The subjects were shown a horizontal line on a computer screen with a vertical slash located in the middle of the horizontal line. Underneath the right end of the horizontal line was the printed word *bow*, and underneath the left end was the printed word *pluck*. The subjects could move a second vertical slash along the horizontal line by manipulating a computer mouse. Upon each presentation, the subjects performed judgments by moving the vertical slash to a position on the horizontal line that corresponded to their impression of the audio token. If the audio token sounded like a clear pluck, the subjects were told that they should place the slash near the left end of the horizontal line. If the audio token sounded like a clear bow, the subjects were to place the slash near the right end of the horizontal line. Finally, if the audio token sounded somewhat ambiguous, the subjects were to place the slash somewhere in the middle of the horizontal line. With the use of this method of graphic rating, identification judgments were quantified in terms of distance of the slash mark from the left end of the line. When the slash mark was placed at the left end of the horizontal line, a value of 0 cm was recorded; when the slash mark was placed at the right end of the line, a value of 15.14 cm was recorded. The subjects made all of their judgments on a Macintosh computer.

After making each identification judgment, the subjects were instructed to rate the discrepancy between the audio and visual components. The purpose of this task was primarily to ensure that the subjects were attending to the video component of the tokens. Previous research has demonstrated that a task of this sort does not interfere with identification judgments (Rosenblum & Fowler, 1991). For the discrepancy ratings, a row of 11 numbers appeared on the screen after the subjects had identified the audio token. The subjects were instructed to choose (1) a number between 5 and 1 on the left end of the row if the audio token sounded like a bow relative to a video pluck; (2) "0" if the audio and visual components of the token were consistent; or (3) a number between 1 and 5 on the right end of the continuum if the audio token sounded like a pluck relative to a video bow. The number chosen between 1 and 5 was to reflect the degree of discrepancy, with a 5 indicating a *very discrepant* token and a 1 indicating a *slightly discrepant* token.

Before the experiment began, the subjects were told to base their identification judgments only on what they heard (Summerfield & McGrath, 1984). However, they were also told that they would need to pay careful attention to the video component of the token in order to make their discrepancy judgments. For the audio-alone trials, the subjects were told to base their identification judgments on what they heard and simply to choose "0" for the discrepancy rating.

After presentation of the instructions, the subjects were presented with four practice trials to familiarize them with the stimuli and the task. The first two practice trials involved a consistent audiovisual pluck token and a discrepant video pluck token (clear audio bow-video pluck). The third and fourth practice trial consisted of a consistent audiovisual bow token and a discrepant video bow token (clear audio pluck-video bow). The experimenter guided the subjects through the practice trials. The subjects were told which video and audio stimuli were being presented, so they were completely aware of the dubbing procedure. The subjects then performed judgments on the 120 presentations. The order of presentation of the six blocks was counterbalanced across subjects. Five subjects were presented with the blocks in the following order: one block audio alone, four blocks audiovisual, and one block audio alone. Three subjects were presented with the blocks in the order: two blocks audiovisual, one block audio alone, two blocks audiovisual, and one block audio alone. Finally, 5 subjects were presented with the

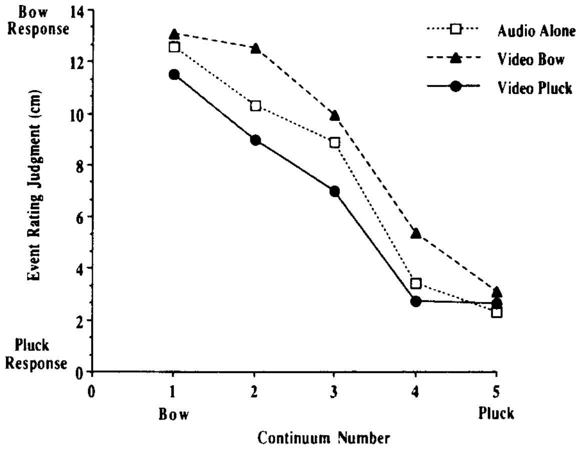

Figure 1. Mean graphic rating of stimulus tokens for audio-alone, video-bow, and video-pluck conditions in Experiment 1.

blocks in the following order: one block audiovisual, one block audio alone, one block audiovisual, one block audio alone, two blocks audiovisual. The entire experiment lasted about 1 h for each subject.

Results and Discussion

The audio-alone judgments were analyzed first and are portrayed graphically in Figure 1. The mean ratings and standard deviations across the 13 subjects in the audio-alone condition can be seen in Table 1. An analysis of variance (ANOVA) showed that there was an overall effect of audio stimulus [$F(4,40) = 24.791, p < .0001$]. This effect indicates that, overall, the subjects were good at differentiating among the five audio tokens.

The mean ratings for the audiovisual conditions are also displayed graphically in Figure 1. These mean ratings, along with their standard deviations, are listed in Table 1. An omnibus ANOVA was performed on the variables of audio level (5), video level (2), and presentation ordering (3). The mean rating for audio tokens presented with a video pluck was 6.556 ($SD = 4.69$); the mean rating for audio tokens presented with a video bow was 8.806 ($SD = 4.94$). This effect of video was significant at the .05 level [$F(1,40) = 8.339$]. This effect indicates that, overall, the subjects' auditory judgments were influenced in the direction of the paired video token. There was also a significant audio × video interaction [$F(4,40) = 2.966, p < .05$] (see Figure 1). A series of planned comparisons revealed a significant effect of video on all levels of audio except for the fifth level [Level 1, $F(1,40) = 8.927, p < .05$; Level 2, $F(1,40) = 27.031, p < .05$; Level 3, $F(1,40) = 20.672, p < .05$; Level 4, $F(1,40) = 14.266, p < .05$].[1] There was no overall effect of stimulus presentation ordering, nor were there any significant interactions involving ordering.

As stated previously, the primary purpose of the discrepancy ratings was to ensure that the subjects were attending to the video portion of the stimuli. Accordingly, the results will only be summarized descriptively. It should be noted that a negative number indicates that the audio token was judged as a bow relative to a video pluck, whereas a positive number indicates that the audio token was judged as a pluck relative to a video bow. Furthermore, the magnitude of the rating away from 0 (toward 5 or −5) indicates the degree to which the audio and video were judged as discrepant. The ratings were pooled across subjects, and the mean ratings were obtained for each of the 10 audiovisual tokens. These results are plotted in Figure 2, and the mean scores along with their standard deviations are listed in Table 1. These mean discrepancy ratings were highly systematic, indicating that subjects were watching the video monitor and were accurate in assessing audiovisual discrepancy. (It should be noted that the discrepancy ratings within subjects were also highly systematic.)

Overall, the data demonstrate that video information had a significant effect on auditory pluck and bow judgments, even when subjects were explicitly told to base their judgments only on what they heard. Thus, we were successful in demonstrating a visual influence on auditory identification judgments of natural nonspeech stimuli.

The question remains whether the nonspeech visual influence observed in Experiment 1 is comparable to the visual effect demonstrated with speech consonants (see, e.g., McGurk & MacDonald, 1976). As mentioned, the FLMP and direct theories both propose that a strong McGurk-type effect could occur with nonspeech stimuli, but the motor theory would likely propose that a true, striking McGurk effect could occur only with speech stimuli. In one strong version of the McGurk effect, the visual influence can make a clearly defined syllable of one type (e.g., /ba/) sound like a clearly defined syllable of another (/va/) (see, e.g., Rosenblum & Saldaña, 1992). In other words, it is as if continuum endpoints can be visually influenced to sound like their opposite endpoints. Although Experiment 1 demonstrated a significant effect of video

Table 1
Pooled Mean Ratings (With Standard Deviations) for Pluck–Bow Continuum in Experiment 1

							Discrepancy Rating			
Auditory	Audio Alone		Video Pluck		Video Bow		Video Pluck		Video Bow	
Continuum	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
1 (Bow)	12.570	3.72	11.512	4.364	13.091	2.354	−3.82	1.25	.01	1.645
2	10.301	3.59	8.986	3.371	12.452	2.473	−2.78	1.426	.80	.917
3	8.888	2.54	6.961	3.366	9.943	3.255	−2.07	1.473	1.425	1.196
4	3.419	2.55	2.704	2.170	5.363	3.515	−.511	.534	3.08	1.23
5 (Pluck)	2.307	2.20	2.617	2.250	3.088	3.324	−.905	1.054	3.52	1.373

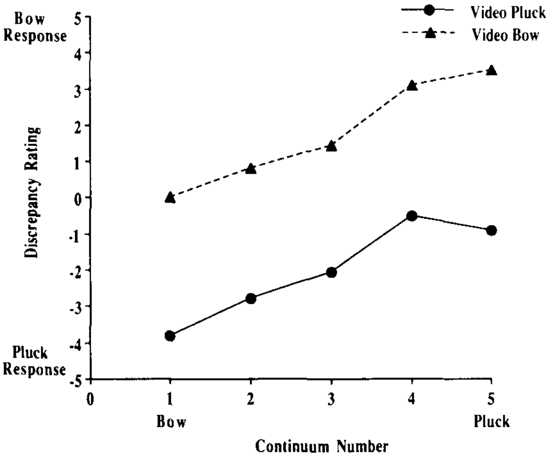

Figure 2. Mean discrepancy ratings for 10 audiovisual tokens presented in Experiment 1.

on subjects' pluck-bow identification judgments, it was not the case that a clear pluck could be made to sound like a clear bow (and vice versa) due to visual influence. Furthermore, the discrepancy rating data suggest that the subjects could distinguish the auditory and visual information (see Figure 2). This would seem quite different from the speech McGurk effect, which is thought to involve percepts that cannot be parsed into auditory and visual components (Liberman & Mattingly, 1985). It would seem, then, that our nonspeech effect does not match the strength of the classic speech McGurk effect.

It has been suggested by a reviewer, however, that the difference between the results of Experiment 1 and those found with speech consonants might have to do with a difference in task rather than a difference in effect strength.^{2} In studies involving visually influenced speech consonants, observers are typically asked simply to identify—rather than rate—the syllable that is heard. Results from an identification task of this sort might erroneously suggest that the visual influence was quite large.

To more directly compare the results of Experiment 1 with results involving speech stimuli, a second experiment was conducted. It involved a rating task for syllable judgments, rather than an identification task.

EXPERIMENT 2

In Experiment 2, subjects were asked to rate their perception of audiovisual syllables. The syllables used in this experiment were /ba/s and /va/s. These syllables were chosen for two reasons. First, these stimuli have been shown to be easily influenced by visual articulatory information (Repp et al., 1983; Rosenblum & Saldaña, 1992). Second, an editing procedure similar to that used in transforming the auditory bow into an auditory pluck could be used to transform the syllable /va/ into a /ba/. This procedure involved cutting off the rising portion of a /va/ token in equal increments so that, again, a five-point continuum could be generated.

As mentioned, the goal of Experiment 2 was to compare more directly the visual influences of consonants with our nonspeech results. Accordingly, measures were taken to ensure that the auditory speech continuum stimuli did not have a greater chance of being visually influenced simply by virtue of being more ambiguous. To ensure that the syllables were not more or less ambiguous than the pluck-bow tokens, speech stimuli were chosen that produced ranges of auditory identification values similar to those for the pluck-bow stimuli.

In the selection of stimuli for Experiment 2, a pilot experiment was conducted to ensure that the endpoint tokens of the speech continuum were rated (roughly) as extreme as were the endpoints of the pluck-bow continuum. Five naive subjects judged a series of audio /ba/-/va/ continuum tokens on a rating scale. This scale was configured so that the left end of the continuum (recorded value of 0 cm) represented a clear /ba/ and the right end (15.14 cm) represented a clear /va/. Results of the pilot experiment revealed that the mean rating for the speech syllable with a 5-msec rising portion was 2.09 (/ba/), which compared most closely with the mean rating (2.307) for the endpoint pluck token in the first experiment. The mean rating for the token with a 30-msec rising portion was 13.33 (/va/), which compared most closely with the mean rating for the endpoint bow token (12.57). Accordingly, these two tokens were chosen as endpoints for our syllable continuum. The remaining three members of the continuum were generated by interpolating the rising portion from one endpoint to the other in 6.25-msec increments.

Method

Subjects. The subjects were 13 undergraduates at the University of California, Riverside. All reported having normal hearing and normal or corrected vision. None had participated in Experiment 1.

Stimuli. A PVS350 camcorder and a Shure SM57 microphone were used to record the initial stimulus tape. The actor was seated 5 ft in front of the camera. The camera's focus was centered on the actor's lips. The recorded image consisted of the visage from the bottom of the actor's nose to the bottom of his chin; no background was visible. The actor was recorded articulating the syllables /ba/ and /va/ four times each.

A /va/ token from the video tape was input into a file on the computer. The computer was then used to generate a five-point continuum based on the outcome of the pilot experiment. The endpoint /va/ token was about 595 msec in duration. With a speech analysis software package (CSRE), a continuum ranging from /va/ to /ba/ was generated by cutting off the rising portion of the endpoint /va/ token in (roughly) 6.25-msec increments (care was taken to edit the signal at zero crossings). Because of the editing procedure, the continuum members differed slightly in duration, with the endpoint /ba/ token being 25 msec shorter than the endpoint /va/ token. This difference was not noticeable to the experimenters.

The audiovisual presentation tape was prepared as in Experiment 1. The number of blocks and the order of the tokens were analogous to those in the first experiment. The presentation ordering was also the same, with 4 subjects participating in Order 1, 4 in Order 2, and 5 in Order 3.

Procedure. The procedure from Experiment 1 was also used in this experiment. However, for this experiment, the rating scale was labeled with "ba" underneath the left end of the horizontal line and "va" underneath the right end of the horizontal line. Again,

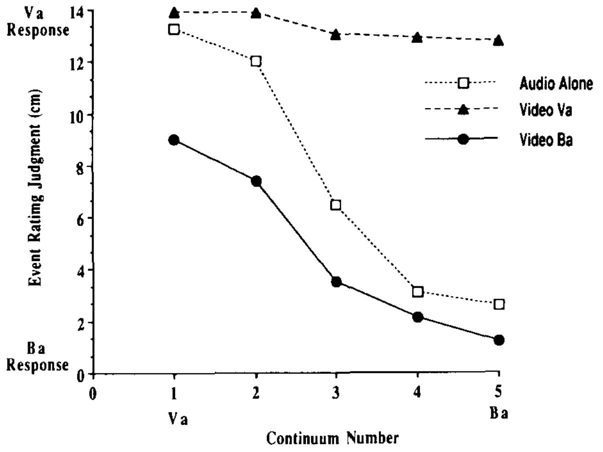

Figure 3. Mean graphic ratings of stimulus tokens for audio-alone, video /va/, and video /ba/ conditions in Experiment 2.

the subjects were told to base their judgments only on what they heard. They also performed discrepancy ratings, to ensure that they watched the video throughout the experiment. The subjects performed judgments on 80 audiovisual tokens and 40 audio-alone tokens. Four practice trials were presented to familiarize the subjects with the task.

Results and Discussion

The audio-alone judgments were analyzed first. The mean ratings across the 13 subjects are portrayed graphically in Figure 3. These means, along with their standard deviations, are listed in Table 2. An ANOVA revealed that there was an overall effect of audio stimulus [$F(4,40) = 68.0, p < .001$]. This indicated that, in general, the subjects were good at differentiating among the five tokens.

Regarding the audiovisual stimuli, the mean ratings for the 10 tokens are also depicted graphically in Figure 3. In addition, these mean ratings along with their standard deviations can be seen in Table 2. The mean rating for audio tokens presented with a video /ba/ was 4.65 ($SD = 3.65$); the mean rating for audio tokens presented with a video /va/ was 13.27 ($SD = 1.54$). An omnibus ANOVA involving the variables of video level, audio level, and presentation ordering revealed a significant main effect of video. This effect of video was highly significant at the .001 level [$F(1,40) = 170$], which suggests that the observers' auditory judgments were visually influenced toward the direction of the visual syllable presented. There was also a significant audio × video interaction [$F(4,40) = 34.0, p < .001$] (see Figure 3), which indicates that some auditory tokens were influenced more strongly by the visual information than others. However, a series of planned comparisons revealed a significant effect of video on all levels of audio [Token 1, $F(1,40) = 97.46, p < .001$; Token 2, $F(1,40) = 166.62, p < .001$; Token 3, $F(1,40) = 365.79$, $p < .001$; Token 4, $F(1,40) = 472.49$, $p < .001$; Token 5, $F(1,40) = 548.69$, $p < .001$].

A final ANOVA was performed that included the variable of experiment (Experiments 1 and 2). This analysis allowed us to compare the influence of video in the nonspeech and speech results. For this analysis, the data were pooled over presentation ordering. The ANOVA revealed an overall effect of video [$F(1,24) = 123.017, p < .001$], as well as a significant video × experiment interaction [$F(1,24) = 42.27, p < .001$]. This interaction indicates that the visual effect demonstrated in Experiment 2 was significantly greater than the visual effect found in Experiment 1. This is easily observable when one compares Figures 1 and 3.

As in Experiment 1, the discrepancy ratings were pooled across subjects to obtain mean discrepancy ratings for each audiovisual token. These data are plotted in Figure 4; the mean discrepancy ratings and standard deviations are listed in Table 2. It should be noted that the discrepancy ratings for the speech stimuli were much smaller than those found in the previous experiment. This result, however, does not mean that the subjects failed to consistently watch the video portion of the presentations. In fact, the degree of visual influence found in Experiment 2 attests to the subjects' attention to the video display. Rather, these smaller and less systematic discrepancy ratings might indicate that the subjects found parsing the auditory and visual portions of these stimuli more difficult than they found parsing the audio and visual portions of the pluck-bow stimuli. This finding would seem to support the claim that the McGurk effect involves percepts whose auditory and visual contributions cannot be distinguished (Liberman & Mattingly, 1985). The fact that a different pattern of discrepancy ratings was found for the speech versus nonspeech stimuli would likely be interpreted by the motor theorist as evidence that our non-

Table 2
Pooled Mean Ratings for /ba/–/va/ Continuum in Experiment 2

							Discrepancy Rating			
Auditory	Audio Alone		Video /ba/		Video /va/		Video /ba/		Video /va/	
Continuum	*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*
1 /va/	13.267	1.335	8.99	2.16	13.89	1.12	−.423	1.656	.606	.579
2	12.014	2.184	7.40	2.58	13.84	1.21	−.442	1.284	.509	.565
3	6.472	3.285	3.51	2.38	13.02	1.89	−.577	1.076	.683	.806
4	3.098	2.24	2.13	1.89	12.88	1.81	−.635	.548	.596	.839
5 /ba/	2.610	1.708	1.21	.84	12.74	1.32	−.510	.728	.529	.807

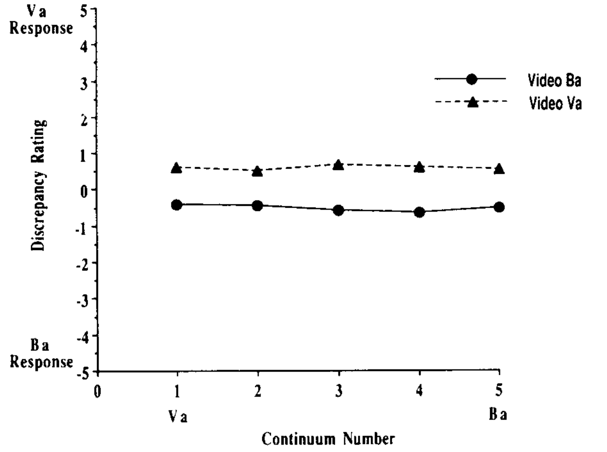

Figure 4. Mean discrepancy ratings for 10 audiovisual tokens presented in Experiment 2.

speech finding did not originate from the mechanism that is responsible for speech McGurk effects.

The identification results confirm our earlier intuition that the visual influence is substantially greater for speech syllables than for our nonspeech stimuli. This finding would seem to be necessary as support for the motor theory, but it is not inconsistent with the direct and FLMP models either. The possible basis for this speech-nonspeech difference will be addressed in the General Discussion.

With regard to the original finding in Experiment 1, why a significant nonspeech visual influence occurred remains an open question. Whereas the motor theory is chiefly concerned with accounting for speech effects, the direct approach and the FLMP would seem to offer two distinct interpretations for this observation. The direct realists propose that a visual influence with either speech or nonspeech is due to a (an ostensibly) lawful relationship between the event and the structured media it produces (Fowler & Dekle, 1991; Fowler & Rosenblum, 1991). On the other hand, the FLMP states that associations of audiovisual cues are built up through experience, and this association is responsible for a visual influence on auditory percepts (Massaro, 1987). The visual influence observed in Experiment 1 could be used as evidence for either of these two assertions, because the visual information presented was both lawfully tied to and associated with the events of plucking and bowing a cello.

Experiment 3 was designed to help determine whether there would be a visual influence of stimuli in which the visual information presented was simply associated with the event by convention. In Experiment 3, the words *pluck* or *bow* were presented in place of the video of the actor playing the cello. If, as suggested by the direct realists, the McGurk effect is due to a (an ostensibly) lawful relationship between the event and the structured media it produces, then text stimuli of this sort—which are only related to the events through convention—should not influence subjects' auditory judgments of the tokens (e.g., Fowler & Dekle, 1991). If, however, the visual influence is due to learned associations, these text stimuli could be effective at influencing subjects' judgments (Massaro, 1987).

In a recent study, Massaro, Cohen, and Thompson (1988) found a small but significant effect of the text "ba" and "da" on "heard" tokens of an auditory /ba/-/da/ continuum (but see Fowler & Deckle, 1991, for contrasting results with text). Although this effect was not nearly as strong as it was when lip-read information was used, it was significant, and it displayed a characteristic patterning of greater influence on more ambiguous auditory tokens. This result is compatible with the FLMP, which predicts no qualitative difference between lip-read and text conditions, with the magnitude of the effect based on the experience of co-occurrence of the auditory and visual events (Massaro et al., 1988). By extension, then, it would seem that the FLMP would predict some slight effect for the words *pluck* and *bow* on auditory judgments. Since it is probable that subjects have more often witnessed the co-occurrence of manual plucking and bowing with pluck and bow sounds than the co-occurrence of the words *pluck* and *bow* with these sounds, the text effect should not be nearly as large. However, according to the FLMP, the text condition should still show a similar patterning of results with a trend of greater visual influence on more ambiguous auditory tokens (see note 1).

EXPERIMENT 3

Method

Subjects. The subjects were 14 undergraduates at the University of California, Riverside. All reported having normal hearing and normal or corrected vision. None had participated in Experiments 1 and 2. Again, these new subjects were not selected for their musical experience (Cutting & Rosner, 1974).

Stimuli. The audio stimuli were the same as in Experiment 1. The video stimuli consisted of the black printed words "Pluck" and "Bow" on a white background. Each of these words was videotaped and then edited onto a new presentation tape. This presentation tape was produced with the methods discussed in Experiments 1 and 2. However, on this tape, the word *pluck* was edited onto any token where the actor plucked the string, and the word *bow* was edited onto any token where the actor bowed the string. The printed words remained on the screen for 3 sec (as had the video events used in Experiment 1). The three presentation orderings were the same as in Experiments 1 and 2, with 6 subjects in the first ordering, 5 subjects in the second ordering, and 3 subjects in the third ordering.

Procedure. The procedure used in Experiments 1 and 2 was again used in this experiment. Again, the subjects were told to base their judgments only on what they heard. They also performed discrepancy ratings to ensure that they watched the video throughout the experiment. The subjects again performed judgments on 80 audiovisual tokens and 40 audio-alone tokens. Four analogous practice trials were implemented at the beginning of the experiment.

Results and Discussion

The audio-alone conditions were analyzed first. The mean ratings across the 14 subjects over the five audio conditions are portrayed in Figure 5. The mean and standard deviation values for the five conditions can also be seen in Table 3. As expected, an ANOVA revealed a sig-

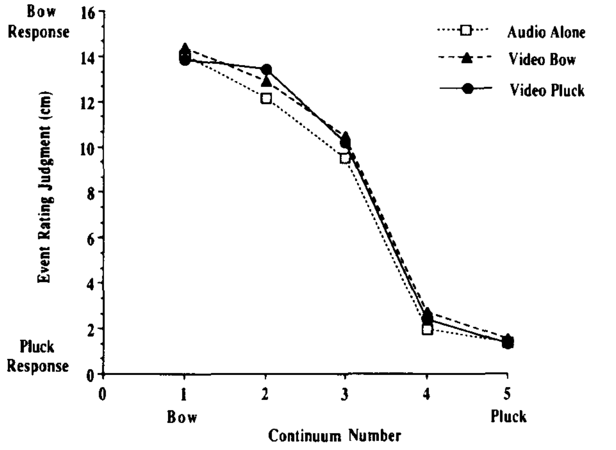

Figure 5. Mean graphic rating of stimulus tokens for audio-alone, text-bow, and text-pluck conditions in Experiment 3.

nificant effect of audio on the subjects' responses [$F(4,44) = 173.9, p < .001$]. As in Experiment 1, it appears that the subjects were able to identify the audio tokens accurately. These results further suggest that the subjects had an established association between the words *pluck* and *bow* with their respective sounds. Recall that in order to rate the auditory tokens as pluck-like or bow-like, the subjects used these words as endpoint anchors in positioning the cursor for each judgment.

We next checked to see whether there was a significant effect of video on what subjects reported "hearing." These results are shown graphically in Figure 5. The mean ratings and standard deviations for the audiovisual tokens are listed in Table 3. The mean rating across all audio tokens when the word *pluck* was presented was 8.22 ($SD = 5.7$); the mean rating across all audio tokens when the word *bow* was presented was 8.4 ($SD = 5.8$). An ANOVA showed no significant effect of video (text) [$F(1,44) = .965, p > .05$], which indicates that the visual information was not successful in influencing the subjects' judgments. There was no effect of presentation ordering [$F(2,44) = .089, p > .05$].

As in the previous experiment, the discrepancy ratings were pooled across subjects to obtain mean ratings for each audiovisual token. This data is portrayed in Figure 6; the means and standard deviations for these judgments can be seen in Table 3. Again, these systematic ratings indicate that the subjects were attending to the video display and were accurate in assessing audiovisual discrepancy (the discrepancy ratings within subjects were highly systematic).

Thus, no influence of the words *pluck* and *bow* on auditory judgments was observed. These findings might suggest that an association by convention is not sufficient for producing a cross-modal perceptual effect. These findings also support the direct realist's contention that a (an ostensive) lawful relationship must exist between the stimuli and event in order for a visual influence to be effective.

How do these results bear on the FLMP? Clearly, subjects have some association between the words *pluck* and *bow* and the sounds of plucks and bows. This is evident from the fact that the subjects in Experiments 1 and 3 had no trouble using the rating scale that had *pluck* and *bow* labels. It would seem that the FLMP would predict that text stimuli should also influence—to some small extent—what subjects hear (Massaro et al., 1988). This prediction was not borne out in Experiment 3. Furthermore, there was no evidence of a characteristic patterning of greater influence on more ambiguous auditory tokens. It is possible, however, that although the subjects' associations between the text and auditory stimuli could allow them to use *pluck* and *bow* labels, these associations were simply not strong enough to produce a significant visual influence in Experiment 3. Thus, although the null results of the text experiment might not necessarily be predicted from the FLMP, these results do not completely contradict the assumptions of the FLMP. In general, however, it would seem that in both speech and nonspeech domains, the sight of an actual event induces a greater visual influence than does representative text.

The findings of Experiment 3 might also suggest that the cross-modal influence reported in Experiment 1 was not simply based on a conscious decision strategy of the subjects (see Rosenblum & Fowler, 1991, for a discussion of this issue). Since the task and instructions were exactly the same in Experiments 1 and 3, it can be assumed that the subjects used the same general strategies in performing their judgments in both experiments. Accordingly, if a conscious decision strategy were used, it

Table 3
Pooled Mean Ratings for Pluck-Bow Continuum in Experiment 3

							Discrepancy Rating			
Auditory	Audio Alone		Video Pluck		Video Bow		Video Pluck		Video Bow	
Continuum	*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*
1 (Bow)	14.023	1.081	13.83	1.405	14.344	.958	−4.14	.931	.429	1.112
2	12.142	1.830	13.411	1.7	12.905	2.127	−3.15	2.175	.59	1.309
3	9.496	1.875	10.177	2.162	10.469	3.027	−2.65	1.018	1.24	1.145
4	1.956	1.875	2.388	2.691	2.71	3.019	−.779	1.216	3.39	1.432
5 (Pluck)	1.391	1.431	1.312	1.597	1.312	1.597	−.4	.854	3.95	1.006

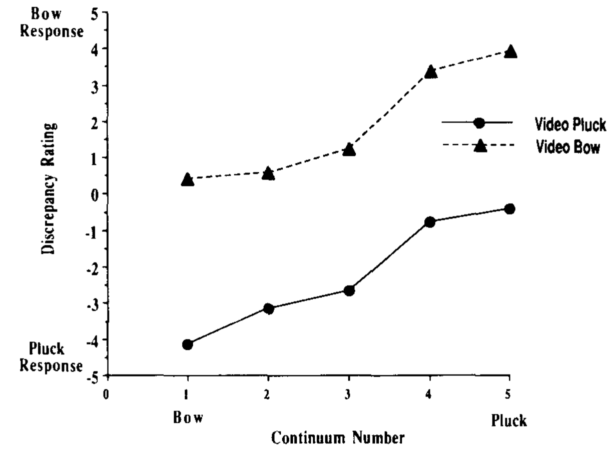

Figure 6. Mean discrepancy ratings for 10 audiovisual tokens presented in Experiment 3.

would be predicted that both experiments would show a visual influence. Since no influence of text on auditory pluck-bow judgments was found, it is probable that the influence of visual information in the first experiment did not have a conscious decision basis.

GENERAL DISCUSSION

Our findings suggest that a visual influence on auditory identification responses can occur with nonspeech stimuli (Experiment 1). At the same time, however, this nonspeech effect is significantly less striking than speech effects (Experiment 2). Why might our nonspeech effects be weaker than those for speech? One answer lies with the motor theorist's contention that a true McGurk effect is a product of a specialized modular process. Again, the motor theorists propose that only for speech is the perceptual object neither visual nor auditory but based on the distal event itself. It follows, then, that striking intermodal influences for which the auditory and visual information cannot be parsed would occur only with speech. Regarding our pluck-bow results, the motor theorist would likely suggest that the more subtle nonspeech visual influence was based in a higher, extramodular, cognitive stage of processing.

However, other interpretations of our findings can be offered that do not proffer a fundamental difference between speech and nonspeech processing. Below, we outline two more interpretations, both of which also address visual influence differences observed for speech vowels as opposed to consonants.

It would seem that our nonspeech results are more similar to McGurk-type effects found for speech vowels than to those found for speech consonants (Cohen & Massaro, 1993; Summerfield & McGrath, 1984). Summerfield and McGrath (1984) found that identifications of tokens along a vowel continuum could be influenced by discrepant visual vowels. This influence, however, was more subtle than that found for consonants: only (significant) shifts in the vowel identification functions were observed. Clearly, these results are more closely analogous to our nonspeech findings. Thus, in order to determine why a difference might exist between speech (consonant) and (our) nonspeech McGurk effects, an understanding of the perceptual differences between consonants and vowels might prove enlightening.

There has been some evidence that consonants are categorically perceived (see Repp, 1984, for a review). In demonstrations of categorical perception, subjects are asked to identify up to 10 synthesized consonant-vowel (CV) tokens belonging to a continuum and differing from each other in equal increments of an acoustic parameter. The subjects usually report hearing only two to three distinct categories of such syllables, and they rarely report hearing any of the tokens as being in between these categories (Liberman et al., 1967). Furthermore, subjects can show a decreased ability to discriminate between tokens that belong to the same category and an increased ability to discriminate between tokens that belong to different categories (Liberman, Harris, Hoffman, & Griffith 1957; Schouten, 1992; but see also Kewley-Port, Watson, & Foyle, 1988; and Watson & Kewley-Port, 1988). The importance of categorical perception for our discussion is that a relatively small change in an acoustic parameter can lead to a rather large perceptual difference for an observer.

The perception of steady-state vowels, on the other hand, appears to be more continuous than categorical (Liberman et al., 1967; but see Pisoni, 1973). For example, if a synthetic continuum ranges from an /i/ to an /I/, subjects report hearing many intraphonemic variations across the continuum. For vowels, then, it is not the case that a relatively small acoustic change can lead to a disproportionately large perceptual change, as is the case for consonants (see Liberman et al., 1967; and Ades, 1977, for possible explanations of the categorical nature of consonants relative to vowels).

Perhaps phenomenally striking "consonant" McGurk effects are due, in part, to the categorical nature of consonants. As mentioned, although a relatively small acoustic change can potentially lead to a large perceptual change for CV syllables, an analogous acoustic change for steady-state vowels will not lead to quantal changes. In principle, it could be that the visual influence for consonant and vowel effects is actually the same in the sense that it pushes the heard token away from the audio-alone percept to the same degree. However, although the visual influence for a consonant might push the auditory percept into another category, a visual influence on a vowel will not have this effect. This would lead to a more phenomenally striking effect for consonants.

Returning to our pluck-bow McGurk effects, it is generally believed that pluck and bow continua tokens are not perceived in a categorical fashion (e.g., Rosen & Howell, 1981; Smurzyński, 1985; but see also Cutting, 1982; Cutting & Rosner, 1974; Cutting, Rosner, & Foard, 1976).

In light of the arguments outlined above, this could explain why the obtained visual influence for these nonspeech stimuli seems more similar to visual effects for vowels than for consonants—the visual influence fails to push the auditory percept into another clear (pluck or bow) category.

It should be noted that not all theorists agree that the difference in visual influence between consonants and vowels is to be found in the categorical as opposed to noncategorical nature of the percept. Proponents of the FLMP (Cohen & Massaro, 1993) argue that both consonants and vowels are perceived continuously. According to the FLMP, the difference observed for vowel and consonant identification functions is largely based on the transient nature of consonants' acoustic signal (Studdert-Kennedy, 1976). On the basis of this difference, it is thought that the auditory information for vowels is much more robust than the auditory information for (at least) stop consonants. According to the FLMP, when subjects identify vowels, they are able to draw on characteristics of the auditory signal itself to make accurate judgments. Conversely, when asked to identify consonants, subjects are less likely to rely on the highly transient auditory signal and instead rely on their *postperceptual* phonetic categorization of the stimuli. This would account for the more "categorical" nature of the identification functions for CV syllable continua.

Regarding audiovisual speech perception, the FLMP suggests that the relative ambiguity of vowel and consonant acoustic signals contributes to the strength of the visual influence. According to the FLMP, there is a tradeoff between audio and visual information: the contribution of one information source grows weaker as the other source becomes less ambiguous. For audiovisual vowels, the auditory information is more robust than the visual information. This is due to the less transient nature of the auditory information, as well as to the fact that visible vowels involve less specific articulator positions and might therefore be less discriminable (Cohen & Massaro, 1993). (It is easier to articulate the same vowel with different vocal tract configurations than it is for most consonants; see Ladefoged, Harshman, Goldstein, & Rice, 1978.) This contrasts with audiovisual consonants for which the auditory information is less robust and the visual information involves more specific visible articulatory positions. According to the FLMP, these differences account for the greater visual influences on consonants than on vowels.

Regarding our nonspeech findings, it is likely that the pluck-bow visual information was at least as robust as the visual information for the /b/-/v/ consonants. Recall that the actor's arm entered the video empty for all video pluck tokens, and with a bow for all video bow tokens. Why, then, did we not find a phenomenally striking effect? The FLMP proponents might suggest that the auditory information for plucks and bows is highly robust in comparison with the auditory information for consonants. Because of the assumption of a tradeoff between the influence of auditory and visual information, this assertion would allow the FLMP to handle the difference between pluck-bow and consonant findings.

In conclusion, the difference in strength for the visual influence of consonants and nonspeech might be based on principles distinguishing visual influences of consonants and vowels. If this is the case, future research should be designed to implement nonspeech sounds that have characteristics of consonants (e.g., categorical perception and/or more transient acoustic properties), in order to demonstrate a more phenomenally striking nonspeech McGurk effect.

REFERENCES

Ades, A. E. (1977). Vowels, consonants, speech and non-speech. *Psychological Review*, **84**, 524-530.

Bermant, R. I., & Welch, R. B. (1976). The effect of degree of visual-auditory stimulus separation and eye position upon the spatial interaction of vision and audition. *Perception & Motor Skills*, **43**, 487-493.

Choe, C. S., Welch, R. B., Guilford, R. M., & Juola, J. F. (1975). The "ventriloquist effect": Visual dominance or response bias? *Perception & Psychophysics*, **18**, 55-60.

Cohen, M. M., & Massaro, D. W. (1993). *Perceiving visual and auditory information in consonant-vowel and vowel syllables*. Manuscript submitted for publication.

Cutting, J. E. (1982). Plucks and bows are categorically perceived, sometimes. *Perception & Psychophysics*, **31**, 462-476.

Cutting, J. E., & Rosner, B. S. (1974). Categories and boundaries in speech and music. *Perception & Psychophysics*, **16**, 564-570.

Cutting, J. E., Rosner, B. S., & Foard, C. F. (1976). Perceptual categories for musiclike sounds: Implications for theories of speech perception. *Quarterly Journal of Experimental Psychology*, **28**, 361-378.

Easton, R. D., & Moran, P. W. A. (1978). A quantitative confirmation of visual capture of curvature: Implications for a visual dominance phenomenon. *Journal of General Psychology*, **98**, 105-112.

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

Fowler, C. A., & 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.

Fowler, C. A., & Rosenblum, L. D. (1991). Perception of the phonetic gesture. In I. G. Mattingly & M. Studdert-Kennedy (Eds.), *Modularity and the motor theory* (pp. 33-59). Hillsdale, NJ: Erlbaum.

Gibson, J. J. (1979). *The senses considered as perceptual systems*. Boston: Houghton Mifflin.

Kewley-Port, D., Watson, C. S., & Foyle, D. C. (1988). Auditory temporal acuity in relation to category boundaries: Speech and nonspeech stimuli. *Journal of the Acoustical Society of America*, **83**, 1133-1145.

Kinney, J. A. S., & Luria, S. M. (1970). Conflicting visual and tactual-kinesthetic stimulation. *Perception & Psychophysics*, **8**, 189-192.

Ladefoged, P., Harshman, R. Goldstein, L., & Rice, B. (1978). Generating vocal tract shapes from formant frequencies. *Journal of the Acoustical Society of America*, **64**, 1027-1035.

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

Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of the speech sound within and across phoneme boundaries. *Journal of Experimental Psychology*, **54**, 358-368.

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

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

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

Massaro, D. W., & Cohen, M. M. (1983). Evaluation and integration of visual and auditory information in speech perception. *Journal of Experimental Psychology: Human Perception & Performance*, **9**, 753-771.

Massaro, D. W., Cohen, M. M., & Thompson, L. A. (1988). Visible language in speech perception: Lipreading and reading. In R. Campbell (Ed.), *Visible language* (Vol. 22, pp. 8-31). Providence, RI: Rhode Island School of Design.

Mattingly, I. G., & Liberman, A. M. (1990). Speech and other auditory modules. In G. M. Edelman, W. E. Gall, & W. M. Cowan (Eds.), *Signal and sense: Local and global order in perceptual maps* (pp. 501-520). New York: Wiley.

McGrath, M., & Summerfield, A. Q. (1985). Intermodal timing relations and audio-visual speech recognition by normal-hearing adults. *Journal of the Acoustical Society of America*, **77**, 678-685.

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

Michaels, C. F., & Carello, C. (1981). *Direct perception*. Englewood Cliffs, NJ: Prentice-Hall.

O'Leary, A., & Rhodes, G. (1984). Cross-modal effects on visual and auditory object perception. *Perception & Psychophysics*, **35**, 565-569.

Pisoni, D. B. (1973). Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Perception & Psychophysics*, **13**, 253-260.

Repp, B. H. (1984). Categorical perception: Issues, methods, findings. In N. J. Lass (Ed.), *Speech and language: Advances in basic research and practice* (Vol. 10, pp. 2-36). New York: Academic Press.

Repp, B. H., Manuel, S. Y., Liberman, A. M., & Studdert-Kennedy, M. (1983). *Exploring the "McGurk effect."* Paper presented at the meeting of the Psychonomic Society, San Diego.

Rosen, S. M., & Howell, P. (1981). Plucks and bows are not categorically perceived. *Perception & Psychophysics*, **30**, 156-168.

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

Rosenblum, L. D., & Fowler, C. A. (1991). Audiovisual investigation of the loudness-effort effect for speech and nonspeech events. *Journal of Experimental Psychology: Human Perception & Performance*, **17**, 976-985.

Rosenblum, L. D., & Saldaña, H. M. (1992). Discrimination tests of visually influenced syllables. *Perception & Psychophysics*, **52**, 461-473.

Schouten, M. E. H. (1992). Modeling phoneme perception I: Categorical perception. *Journal of the Acoustical Society of America*, **92**, 1841-1855.

Singer, G., & Day, R. H. (1969). Visual capture of haptically judged depth. *Perception & Psychophysics*, **5**, 315-316.

Smurzyński, J. (1985). Noncategorical identification of rise time. *Perception & Psychophysics*, **38**, 540-542.

Studdert-Kennedy, M. (1976). Speech perception. In N. J. Lass (Ed.), *Contemporary issues in experimental phonetics* (pp. 243-293). New York: Academic Press.

Summerfield, A. Q. (1979). Use of visual information in phonetic perception. *Phonetica*, **36**, 314-331.

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.

Walker, J. T., & Scott, K. J. (1981). Auditory-visual conflicts in the perceived duration of lights, tones, and gaps. *Journal of Experimental Psychology: Human Perception & Performance*, **7**, 1327-1339.

Watson, C. S., & Kewley-Port, D. (1988). Some remarks on Pastore. *Journal of the Acoustical Society of America*, **84**, 2266-2270.

Welch, R. B., DuttonHurt, L. D., & Warren, D. H. (1986). Contributions of audition and vision to temporal rate perception. *Perception & Psychophysics*, **39**, 294-300.

NOTES

1. It should be noted that the FLMP predicts that the visual effect should be larger for the more ambiguous auditory stimuli. This prediction is partially borne out in Experiment 1, in that there is a significant effect for the auditory tokens 2, 3, and 4, which are the more ambiguous of the five tokens. However, there is also a small but significant effect of video on the auditory token 1, which is the clearest bow token.

2. We would like to thank Dominic Massaro for suggesting this experiment.

(Manuscript received March 12, 1992;
revision accepted for publication February 19, 1993.)