Perception & Psychophysics
1991, 50 (6), 524-536

Integrating speech information across talkers, gender, and sensory modality: Female faces and male voices in the McGurk effect

KERRY P. GREEN
University of Arizona, Tucson, Arizona

and

PATRICIA K. KUHL, ANDREW N. MELTZOFF, and ERICA B. STEVENS
University of Washington, Seattle, Washington

Studies of the McGurk effect have shown that when discrepant phonetic information is delivered to the auditory and visual modalities, the information is combined into a new percept not originally presented to either modality. In typical experiments, the auditory and visual speech signals are generated by the same talker. The present experiment examined whether a discrepancy in the gender of the talker between the auditory and visual signals would influence the magnitude of the McGurk effect. A male talker's voice was dubbed onto a videotape containing a female talker's face, and vice versa. The gender-incongruent videotapes were compared with gender-congruent videotapes, in which a male talker's voice was dubbed onto a male face and a female talker's voice was dubbed onto a female face. Even though there was a clear incompatibility in talker characteristics between the auditory and visual signals on the incongruent videotapes, the resulting magnitude of the McGurk effect was not significantly different for the incongruent as opposed to the congruent videotapes. The results indicate that the mechanism for integrating speech information from the auditory and the visual modalities is not disrupted by a gender incompatibility even when it is perceptually apparent. The findings are compatible with the theoretical notion that information about voice characteristics of the talker is extracted and used to normalize the speech signal at an early stage of phonetic processing, prior to the integration of the auditory and the visual information.

Over the past four decades, extensive research has been done on the psychological processes underlying the perception and production of spoken language. Much of this research has focused on how the listener processes the acoustic structure of speech in order to arrive at the intended meaning of an utterance.

Although speech perception has primarily been considered an auditory process, recent studies have shown that visual information provided by movements of a talker's mouth and face strongly influences what an observer perceives (Green & Kuhl, 1989, 1991; Green & Miller, 1985; MacDonald & McGurk, 1978; Massaro & Cohen, 1983; McGurk & MacDonald, 1976; Reisberg, McLean, & Goldfield, 1987; Summerfield & McGrath, 1984). A particularly convincing demonstration of the effects of vi sion on speech perception is provided by the stimulus sit uation in which the separate auditory and visual input seem to fuse or blend into a new percept—one that ha not been presented to either modality alone. For exam ple, in the McGurk effect (McGurk & MacDonald, 1976) the auditory syllable /ba/ is presented in synchrony witl a videotape of the talker pronouncing the syllable /ga/ the resulting syllable is perceived as /da/, a syllable tha has not been presented to either modality and that repre sents a combination of both.

This fusion effect is somewhat surprising, because i was long assumed that visual information in the form o lipreading was effective only when the auditory signal wa degraded (Sumby & Pollack, 1954). The McGurk effec demonstrated that visual information is potent even whe the auditory signal is clear and unambiguous. From th standpoint of theory, we have to explain how such divers acoustic and optic information—frequency transitions ir dicating /b/ in the auditory domain and mouth movement indicating /g/ in the visual domain—are combined t produce /d/ by the perceptual system. Although the phe nomenon itself has been well-documented (Green & Kuh 1989, 1991; MacDonald & McGurk, 1978; Manuel, Repp

This research was supported by National Institutes of Health Grant NS-26475 to Kerry P. Green and National Institutes of Health Grant HD-18286 to Patricia K. Kuhl. We would like to thank Virginia Mann and an anonymous reviewer for helpful comments on a previous version of the manuscript. A portion of these data were presented at the spring meeting of the Acoustical Society of America, State College, Pennsylvania, 1990. Correspondence concerning the article should be addressed to Kerry P. Green, Cognitive Science, Psychology Building, Room 312, University of Arizona, Tucson, AZ 85721.

Studdert-Kennedy, & Liberman, 1983; Massaro & Cohen, 1983; McGurk & MacDonald, 1976; Mills & Thiem, 1980; Roberts & Summerfield, 1981), the boundary conditions of the phenomenon and the particular circumstances that affect when the auditory and visual information are integrated remain to be charted (Summerfield, 1987).

It has been suggested (e.g., by Welch & Warren, 1980) that studying how perceptual systems deal with intermodal discrepancies will inform us about the intermodal organization that underlies normal perception. For example, Welch and Warren have proposed a model that describes factors affecting the integration of information in multimodal situations. One important assumption of their model is that of *unity*. The perceiver forms an assumption about whether he or she is observing a single or a multiple event. If the information from the two modalities is perceived as consistent, then a high-unity assumption is produced and the information is treated as belonging to a single event. Under these conditions, the information is combined, even though it is actually discrepant. Alternatively, if the information from the two modalities is perceived as discrepant, a low-unity assumption is produced and the observer treats the information from the two modalities as separate and belonging to two different events. Under these circumstances, the information is not combined.

The ventriloquism effect is an example of a multimodal situation in which the unity assumption holds. When there is a spatial discrepancy between the visual and auditory locations of a sound's source, observers typically hear the sound as emanating from the spatial location of the visual source. Studies of this phenomenon have shown that it is susceptible to a number of factors, including the congruence or *cognitive compellingness* of the auditory and visual information (Jack & Thurlow, 1973; Jackson, 1953; Warren, 1979; Warren, Welch, & McCarthy, 1981). Congruence, or cognitive compellingness, derives from factors such as the temporal congruence between the auditory and visual signals and the extent to which the two streams of information appear to go together.

Regarding temporal congruence, Jack and Thurlow (1973) demonstrated that when a puppet's mouth movements are temporally synchronized with ongoing speech sounds, there is a greater displacement in the perceived localization of the sound than there is when speech is not temporally coincident with the mouth movements. Thus, the temporal synchrony of the auditory and visual information produced a more compelling event, leading to greater displacement of the auditory sound source. Warren et al. (1981) performed a similar experiment by using a videotape of a talker reading a passage. Again, when the auditory signal from the videotape was temporally congruent with the video signal, observers perceived greater displacement of the location of the auditory source than they did when the auditory signal was temporally displaced.

Warren et al. (1981) also demonstrated that cognitive congruency had an influence on the ventriloquism effect. When the video signal of the talker's face was replaced by a dot on the video monitor, there was very little displacement in perceived localization of the sound source. Finally, Jackson (1953) found that the amount of displacement in perceived localization was increased when the characteristics of the auditory signal (a whistle sound) matched the characteristics of the visual signal (a steam kettle with a puff of steam coming out of it vs. a steam kettle with no steam). Taken together, these several studies demonstrate that a reduction in either the temporal or the cognitive congruency between the auditory and visual signals dramatically reduces the magnitude of the ventriloquism effect.

A question that arises is whether the McGurk speech effect, like the ventriloquism effect, is also influenced by the unity or congruency of the auditory and visual signals. The unity question has not been adequately addressed for the speech case. To date, studies of the impact of temporal asynchrony on the McGurk effect have been inconclusive. Cohen (1984) reported that temporal asynchronies of up to 200 msec have little influence on the magnitude of the McGurk effect. However, it is not clear how aware the subjects were of a discrepancy between the auditory and visual signals. Dixon and Spitz (1980) have shown that observers simply may not be able to detect onset asynchronies between auditory and visual speech information for temporal differences of less than 190 msec, and these are similar to the values used by Cohen (1984).[1] Moreover, other studies indicate that temporal asynchronies of 80–400 msec can disrupt the integration of auditory and visual speech information under some circumstances in the McGurk situation (McGrath & Summerfield,1985) and other situations as well (Dodd, 1977, 1979). Thus, it remains unclear from these studies whether, or how much, the McGurk effect is affected by a reduction in the temporal congruence between the auditory and visual signals.

No studies have directly examined whether changes in the cognitive congruency of the auditory and visual information alter the McGurk effect. The specific purpose of the present study was to manipulate the cognitive congruence between the auditory and visual signals. This was achieved by having perceivers view a novel combination of auditory and visual information—a gender discrepancy produced by combining a male talker's voice with the video of a female talker's face, and vice versa. In most previous experiments on the McGurk effect, the same talker has produced both the auditory and the visual signals.[2] If the McGurk effect is influenced by the cognitive congruency, and the perceptual "unity" of the signals, there ought to be a weaker McGurk effect in the gender-discrepancy condition, wherein the two streams cannot have been produced by the same person (cf. Welch, 1989). Such a finding would suggest that speech fusion effects are similar to other types of perceptual phenomena—that they are sensitive to intersensory discrepancies and are thus appropriately characterized by models such as that proposed by Welch and Warren (1980).

Research on the perception of speech suggests an alternative possibility, however. This work has been directed at how the perceptual system handles the large

amount of acoustic variation in the realization of phonetic segments spoken by different talkers (for reviews, see J. D. Miller, 1989; Nearey, 1989; Strange, 1989; Syrdal & Gopal, 1986). Current theories hold that the perceptual system somehow "normalizes" the acoustic information with respect to talker differences during speech processing. The theories propose that talker information is extracted at an early, "mandatory" step in speech processing (Mullennix & Pisoni, 1990; Mullennix, Pisoni, & Martin, 1989). This view is supported by studies of young infants, which indicate that such normalization mechanisms may be part of the infant's innate perceptual capacities for perceiving speech (Kuhl, 1980, 1985). For example, at as early an age as 6 months, infants can ignore the variability created by different talkers in the acoustic realization of phonetic segments and attend to just the phonetic similarity of speech tokens (Kuhl, 1979, 1983).

Inasmuch as voice characteristics of the talker are extracted early on in speech processing, it is possible that auditory and visual phonetic information are integrated *after* the auditory signal has been normalized in this way. More specifically, it could be hypothesized that at the time of integration of the phonetic information from the two modalities, the auditory information (and perhaps the visual information as well) would already be talker-neutral rather than talker-specific. By this account, the speech fusion effect would *not* be disrupted by a gender discrepancy between the auditory and visual signals. Even though there is a reduction in the unity of the information due to a discrepancy between the auditory and visual signals, the McGurk effect would be unaffected because the phonetic information that is integrated is neutral with respect to talker differences.

EXPERIMENT 1

The question addressed in Experiment 1 was whether an obvious cross-gender discrepancy (e.g., a female face combined with a male voice) reduces the magnitude of the McGurk effect when compared with the situation in which a talker of the same gender produces both the auditory and the visual signals. The cross-gender combination was constructed by dubbing a male voice onto a female talker's face, and by dubbing a female voice onto a male talker's face. These two incongruent situations were compared with their congruent control situations, in which the male voice was paired with the male face and the female voice was paired with the female face. Moreover, as in previous studies of the McGurk effect (MacDonald & McGurk, 1978; Massaro & Cohen, 1983; McGurk & MacDonald, 1976), the stimuli were produced in an /a/ vowel context. However, vowel context can influence both the acoustic cues associated with place of articulation in stop consonants (Dorman, Studdert-Kennedy, & Raphael, 1977; Fischer-Jorgensen, 1954; Repp & Lin, 1989) and the visual cues (Benguerel & Pichora-Fuller, 1982; Erber, 1971). Moreover, Green, Kuhl, and Meltzoff (1988) have provided evidence that vowel context can influence the magnitude of the McGurk effect. In their study, an /i/ vowel context produced the strongest McGurk effect, an /a/ context produced a moderate effect, and an /u/ context almost no effect. Therefore, in order to ensure that the outcome of the gender discrepancy test was not confined to the specific stimuli or vowel context used, the study also included /i/ vowel stimuli.

Method

Subjects

The subjects were 44 undergraduate students who were either paid or given course credit as an incentive to participate. None of the subjects reported any history of a speech or hearing disorder, and all had normal or corrected-to-normal vision. All were native English speakers.

Materials

Visual stimuli. The visual (V) stimuli were prepared by videotaping a male and a female talker while they produced several instances of the syllables /ba/ and /ga/. A color camera (RCA TK45), a microphone (Sony ECM50), and a ¾-in. videotape recorder (JVC CR8250) were used to record the utterances. The talker was seated on a stool in front of a black background. The camera was centered on the talker's face with lighting provided by two video lights, one placed on either side of the talker. This resulted in an excellent view of the talker's mouth and face. The entire face of the speaker, including the hair, was visible, and the faces were clearly distinguishable as male or female. The size of the facial image when shown on a 13-in. video monitor (NEC JC-1215MA) was approximately 12 cm wide and 15 cm long for the female face; the male face was approximately 9 cm wide and 13 cm long. From these recordings, two syllables were selected for each talker, consisting of single tokens of /ba/ and /ga/. The tokens were selected so that their overall durations were similar and the articulations lacked any extraneous movements. These tokens were then edited onto a new videotape. A video editing console (JVC VE92) was used; it was connected to two ¾-in. videocassette machines (JVC CP5550 and CR8250).

Auditory stimuli. The auditory (A) stimuli consisted of the syllables /ba/ and /ga/ spoken by the male and female talkers. The talkers were recorded with a microphone (ElectroVoice 635A) and tape recorder (Nagra III) while they produced several repetitions of each of the syllables in a soundproof room. The syllables were digitized at a 20-kHz sampling rate on a computer (LSI-11/73), low-pass filtered at 9.89 kHz, and analyzed with a signal processing package. For each talker, a single /ba/ and /ga/ with similar durations, which closely matched the duration of the corresponding video tokens, were selected for the experiment.

Auditory-visual simuli. Four types of auditory-visual (AV) stimuli were created. In two of these, the auditory and visual signal originated from talkers of the same gender. These stimuli were called *congruent*, because the talkers, both auditorially and visually, had the same gender. Two other types of stimuli were created by cross-dubbing the visual and auditory information with respect to gender. These were called *incongruent*, because the talkers, both auditorially and visually, had different genders.

For both the congruent and the incongruent stimulus types, all possible pairings of the auditory and visual /ba/ and /ga/ were created, resulting in four AV stimuli. Two of the four AV stimuli provided conflicting phonetic information (e.g., auditory /ba/ was paired with visual /ga/). This is a stimulus for which subjects typically report perceiving a /da/ or /ða/ syllable. It is a blend of information from both modalities, referred to here as a *fusion* response. The second conflicting AV stimulus paired auditory /ga/ with visual

/ba/. This situation typically produces a /bga/ response, which reflects a combination of the phonetic information presented to both modalities; it will be referred to as a *combination* response. The final two AV stimuli were controls, in that they provided matching phonetic information: auditory /ba/ was paired with visual /ba/ and auditory /ga/ was paired with visual /ga/.

Blocks of AV stimuli were edited onto a test videotape. A block of trials consisted of 10 repetitions of a set of four AV stimuli in random order, for a total of 40 trials. Within each block of trials, only one of the four types of stimuli was presented: female face congruent (FC), male face congruent (MC), female face incongruent (FI), or male face incongruent (MI). In addition, 8 practice trials, consisting of two repetitions of each of the four stimuli, were created at the start of the block of 40 trials.

Individual trials consisted of a single AV stimulus, which was preceded and followed by a 1-sec video display of the talker's face with a neutral expression. In addition, each stimulus was edited with a 1-sec fade-up from video black at its start, and a 1-sec fade-out to video black at its conclusion. The fades prevented abrupt visual onsets and offsets, which could have caused masking or interference. The onset-to-onset trial interval was 6 sec.

The AV stimuli were prepared by combining the audio and video tokens. The A stimuli were dubbed onto the edited videotape by playing the videotape on a videocassette recorder (JVC CR8250). The output of one audio channel from the videotape recorder, which contained the original sound track, was fed into a Schmidt trigger of the 11/73 computer. A marker tone preceded each of the utterances on the original sound track. When the Schmidt trigger sensed the onset of the marker tone, it triggered the computer to output one new audio token according to a predetermined order onto the second audio track of the videotape. The new audio tokens were synchronized to occur with the video articulations by using a temporal delay, which was precalculated so that the release burst of the dubbed utterance would match precisely the release burst of the original utterance corresponding to the video token. The syllables were output at 20 kHz and lowpass filtered at 9.89 kHz. This dubbing procedure resulted in a high degree of accuracy in aligning the auditory and visual components for the AV stimuli. Measurements of the first 10 trials of each videotape indicated that the range of asynchronies between the A and V tokens was from +3 to −4 msec.

In addition to the stimuli described above, an identical set of stimuli was created in the exact same manner, using the audio and video tokens of /bi/ and /gi/ produced by the same two talkers.

Procedure

Three separate conditions were run. In the auditory-visual condition, 24 subjects each participated in a 30-min session. Subjects were randomly assigned to one of two experimental conditions: either the female face (F) condition, in which the subjects were presented with the FC and FI trial blocks for both the /a/ and the /i/ vowel contexts, or the male face (M) condition, in which they were presented with the MC and MI trial blocks for both the /a/ and the /i/ vowel contexts. Thus, each subject was presented with a total of four blocks of 40 test trials. A between-subjects design was chosen, because it had been determined in pilot testing that to present all eight blocks of stimuli (a total of 320 trials) in a single testing session required almost 1 h of testing and was too exhausting for the subject. The subjects were counterbalanced according to whether they received the congruent or incongruent stimuli first and also according to vowel context.

The subjects were instructed to watch and listen to each trial, and to identify whether the syllable they heard began with /b/, /d/, /g/, /ð/ or /bg/ (the instructions are presented in the Appendix). Pretesting had determined that these five responses constituted greater than 96% of subjects' perceptions of these stimuli. Thus, in order to simplify scoring and analysis of the data, the subjects were restricted to these five response categories. None of the subjects indicated at the end of the experiment that they felt restricted in their responses by these five categories. In addition, it was emphasized to each subject that although they were provided with five response categories, they were not obligated to use all five of them. The subjects responded verbally to the experimenter, who recorded this information on an answer sheet. This enabled the subjects to keep their attention and vision focused on the video monitor at all times.

In the visual condition, 10 new subjects were presented with only the visual portion of the AV stimuli. As in the auditory-visual condition, each subject was presented with the practice trials followed by the test trials for each of the two blocks, corresponding to the two vowel contexts. Since just the visual portion of the videotape was presented in this condition, the two visual tokens in each block were presented a total of 20 times. The subjects were provided with the five response categories used in the auditory-visual condition, and they again responded verbally so that they could maintain their attention on the video monitor.

In the auditory condition, ten new subjects were presented with only the auditory portion of the AV stimuli. As in the auditory-visual condition, each subject was presented with the practice trials followed by the test trials for each of the two blocks. Since just the auditory portion of the videotape was presented in this condition, a total of 20 repetitions of each of the two auditory tokens was presented in each block. These subjects were also provided with the same five response categories and responded verbally.

All subjects were tested individually in a small, dimly lit, sound-attenuated room. The subject sat at a desk located approximately 115 cm from the video monitor. The monitor was seated on a table behind a large paper panel with a window cut from it, so that the monitor could be viewed by the subjects. The videotape was played on a videocassette machine located in an adjoining control room. During the auditory-visual session, the audio and video outputs from the videocassette player were presented via the video monitor. During the visual condition, the audio signal was disconnected and only the video signal was presented over the video monitor, whereas during the auditory condition, the video signal was disconnected and the audio signal was presented via the loudspeaker in the video monitor. The contrast and brightness controls were both set at about their midpoint levels, and the audio signal was presented at a comfortable listening level of approximately 65 dB SPL, measured for the peak intensity of the vowel at the approximate location of the subject's head.

Results

Unimodal Results

The auditory tokens were responded to at 98% correct or better, regardless of the talker (male or female) or the vowel context (/a/ or /i/). Thus, the auditory tokens were unambiguous with regard to their phonetic specification. Next, consider the results for the V tokens. The overall percent correct responses for the visual conditions are presented in Table 1. As shown, the /b__/ tokens were much more accurately perceived than the /g__/ tokens for both talkers and in both vowel contexts. Some of this is due to the fact that the velar tokens were sometimes confused with either "d__" or "th__," particularly in the /i/ vowel context. This finding is similar to those from previous studies of the accuracy of lip-read consonants (e.g., Binnie, Montgomery, & Jackson, 1974; Woodward & Barber, 1960). Interestingly, in the /a/ vowel context, the velar tokens were often perceived as "bga" for both

Table 1
Mean Percent Response for the Visual Only Condition for the Female and Male Voices Across Two Vowel Contexts

Vowel Context	Visual Consonant	Response Category /b_/	/d_/	/th_/	/g_/	/bg_/
		Female				
/a/	/b_/	87	2	3	1	7
	/g_/	2	16	10	40	32
/i/	/b_/	68	3	2	1	26
	/g_/	2	30	24	39	5
		Male				
/a/	/b_/	82	1	1	1	15
	/g_/	0	28	5	34	33
/i/	/b_/	71	0	0	0	29
	/g_/	2	23	6	59	10

talkers. This finding is probably due to the fact that both talkers started their articulations with their lips closed. The opening of the lips prior to the articulation of the velar consonant would then be consistent with the utterance of an initial bilabial consonant, leading the subjects to sometimes produce "bg__" as a response. It is unclear why such responses occurred less frequently in the /i/ vowel context, although it might be due to the spreading of the lips during the opening of the mouth for the velar consonant. It is possible that this spreading is less consistent with the perception of a bilabial consonant.

Cross-Modal Results

Tables 2 and 3 present the average response percentages for the AV tokens. The data for the /a/ vowel tokens are given in Table 2; the /i/ vowel data are given in Table 3. Consider first the results for the /a/ tokens in the F condition (Table 2). In the FC condition, the fusion stimuli (auditory /ba/ and visual /ga/) produced very few "ba" responses (i.e., very few subjects reported perceiving the actual A stimulus that was presented). Many of the responses were fusion responses (either "d" or "th," for a total of 50%), although a considerable number (37%) of "g" responses were also given. For the combination stimuli (auditory /ga/ with visual /ba/), the majority of the responses were the combination "bga" responses (60%). This pattern of results is similar to those of previous studies (e.g., MacDonald & McGurk, 1978; Manuel et al., 1983; McGurk & MacDonald, 1976). Next, consider the data for the FI condition. These stimuli also produced a substantial McGurk effect. Most of the responses to the fusion stimuli were again fusion responses (a total of 73%), whereas the responses to the combination stimuli were predominantly combination responses (74%).

The results for the M condition are shown at the bottom of Table 2. As in the F condition, the tokens in the M condition also produced McGurk effects, regardless of whether the auditory and visual signal were congruent or incongruent with respect to the gender of the talker. In the MC condition, the fusion stimuli again produced mostly fusion responses (a total of 91%) and very few "ba" responses (9%), whereas the combination stimuli produced a majority of "bga" responses (67%). The results for the MI condition produced a similar pattern: mostly fusion responses to the fusion stimuli (82%) and a majority of "bga" responses to the combination stimuli (56%).

A comparison of the male and female voices in the F and M conditions indicates little difference in the overall magnitude of the McGurk effect regardless of the face onto which the auditory voice is dubbed. To examine such differences, two separate, two-way analyses of variance (ANOVAs) were conducted. The first analysis was of the percent "ba" responses for the fusion stimuli, whereas the second was of the percent "ga" responses for the com-

Table 2
Mean Percent Response for the Female Face and the Male Face Paired With a Female and a Male Voice in the /a/ Vowel Context

Voice	Auditory Consonant	Visual Consonant	Response Category /b_/	/d_/	/th_/	/g_/	/bg_/
			Female Face				
Female	/b_/	/b_/	100	0	0	0	0
	/b_/	/g_/	12	6*	44*	37	1
	/g_/	/b_/	0	0	0	40	60*
	/g_/	/g_/	0	0	1	99	0
Male	/b_/	/b_/	98	0	1	0	1
	/b_/	/g_/	18	16*	58*	7	0
	/g_/	/b_/	2	0	0	24	74*
	/g_/	/g_/	0	0	0	99	1
			Male Face				
Male	/b_/	/b_/	87	0	12	0	1
	/b_/	/g_/	9	14*	77*	0	0
	/g_/	/b_/	0	0	0	33	67*
	/g_/	/g_/	0	0	0	97	3
Female	/b_/	/b_/	89	0	11	0	0
	/b_/	/g_/	16	19*	63*	2	0
	/g_/	/b_/	6	0	2	36	56*
	/g_/	/g_/	0	0	0	97	3

*Typical fusion or combination responses.

Table 3
Mean Percent Response for the Female Face and the Male Face Paired With a Female and a Male Voice in the /i/ Vowel Context

Voice	Auditory Consonant	Visual Consonant	Response Category /b_/	/d_/	/th_/	/g_/	/bg_/
			Female Face				
Female	/b_/	/b_/	100	0	0	0	0
	/b_/	/g_/	7	77*	16*	0	0
	/g_/	/b_/	0	0	0	37	63*
	/g_/	/g_/	0	0	0	100	0
Male	/b_/	/b_/	100	0	0	0	0
	/b_/	/g_/	17	68*	12*	3	0
	/g_/	/b_/	0	0	0	27	73*
	/g_/	/g_/	2	1	0	97	0
			Male Face				
Male	/b_/	/b_/	99	1	0	0	0
	/b_/	/g_/	25	47*	28*	0	0
	/g_/	/b_/	0	0	0	27	73*
	/g_/	/g_/	0	0	0	94	6
Female	/b_/	/b_/	98	0	1	1	0
	/b_/	/g_/	17	24*	59*	0	0
	/g_/	/b_/	0	0	0	28	72*
	/g_/	/g_/	0	0	0	97	3

*Typical fusion or combination responses.

bination stimuli. For both analyses, the two factors were face (male vs. female) and voice (male vs. female).[3]

The results of the analysis for the fusion stimuli indicated no reliable differences for either the voice [$F(1,22) = .006$, $p > .92$] or the face factors [$F(1,22) = .062$, $p > .83$]. More importantly, there was no reliable interaction between the two factors [$F(1,22) = 1.40$, $p > .25$]; such an interaction would be expected to occur if a cross-dubbing of the two voices dramatically reduced the fusion effect. An ANOVA of the combination stimuli yielded similar results for the face and the face × voice interactions [$F(1,22) = .033$, $p > .85$, and $F(1,22) = 2.11$, $p > .16$, respectively]. However, there was an effect for the voice factor [$F(1,22) = 3.99$, $p = .058$]. This is due to the female voice's producing slightly more "g" responses (and therefore fewer combination responses) than the male voice did, regardless of the face on which it was dubbed.

The results for the /i/ vowel tokens are presented in Table 3. As can be seen, the /i/ vowel tokens produced a similar pattern of results with respect to the congruency between the auditory and visual signals. Two separate ANOVAs of the fusion and combination stimuli again revealed no reliable interaction of the face and voice factors [$F(1,22) = .205$, $p > .65$, and $F(1,22) = .521$, $p > .47$] for the fusion and combination responses, respectively. The main effects of face and voice were also nonsignificant for the combination responses [$F(1,22) = .067$, $p > .79$, and $F(1,22) = .728$, $p > .4$, respectively], while the effect of face was insignificant for the fusion responses [$F(1,22) = .452$, $p > .5$]. There was a significant effect of voice for the fusion responses [$F(1,22) = 10.04$, $p < .005$], although this result was due to the greater number of "b" responses produced by the male voice. Again, this occurred regardless of which face the male voice was dubbed onto. Apparently this particular male /i/ token was just not as successful in producing the McGurk effect as the female token was, although why this should be the case is currently unknown.[4]

In summary, the results from this experiment indicate that the McGurk effect is unaffected by differences in the congruity of the auditory and visual signals with respect to the gender of the talker; the effect is equally strong, regardless of whether face-voice stimuli are gender-compatible or gender-incompatible. However, an additional issue needed to be addressed. It was possible that the faces and voices of the two talkers used in Experiment 1 were not perceived to be discrepant when cross-dubbed onto each other. We addressed this issue in two follow-up experiments.

The first (Experiment 2) involved a replication of Experiment 1 using audio tokens produced by a new male and female talker dubbed onto the visual tokens from Experiment 1. The purpose of this experiment was to determine whether the results of Experiment 1 might be attributable, in part, to the particular faces and voices used in the experiment. The purpose of the second follow-up experiment (Experiment 3) was to assess directly whether the discrepancy between the auditory and visual signals was detected by the subjects. In this experiment, the auditory and visual tokens were presented to a new group of subjects, who were asked to rate the compatibility between the auditory and visual signals.

EXPERIMENT 2

In Experiment 2, the generality of the results of Experiment 1 was examined by using audio tokens from two new talkers dubbed onto the visual stimuli used in Experiment 1.

Method

Subjects

A new group of 34 undergraduate students participated in Experiment 2. Each subject was either paid or given course credit as an incentive to participate. None of the subjects reported any history of a speech or hearing disorder, and all had normal or corrected-to-normal vision. All were native speakers of English.

Materials

A new male and female talker were audio-recorded while they said the syllables /ba/, /ga/, /bi/, and /gi/. These stimuli were recorded, measured, and dubbed onto the visual stimuli from Experiment 1, using the apparatus and procedures from that experiment.

Procedures

These new stimuli were presented to a group of 24 subjects in an auditory-visual condition. In addition, a group of 10 subjects were presented with just the auditory tokens in an auditory condition. The procedures and equipment for presenting the stimuli and collecting the responses were identical to those in Experiment 1.

Results

The results for the auditory condition revealed that the new A tokens were identified with 100% accuracy. The results for the auditory-visual condition are presented in Tables 4 and 5. As can be seen in the tables, the results from this experiment closely replicate the findings of Experiment 1. For example, consider the results for the /a/ tokens (Table 4). In the FC condition, the fusion stimuli again produced a considerable number of fusion ("da" or "tha") responses (76%), while the combination stimuli produced a majority of combination ("bga") responses (72%). The same was true for the FI condition (84% and 75% for the fusion and combination responses, respectively). A similar pattern of responses occurred for the M condition: the fusion stimuli produced a considerable number of fusion responses in the congruent and incongruent conditions (71% and 83%, respectively), while the combination stimuli in the congruent and incongruent conditions produced "bga" responses that were comparable to those obtained in Experiment 1 (63% and 48%, respectively). The /i/ vowel tokens produced a similar pattern of responses regardless of the congruency between the A and V signals (Table 5).

Four two-way ANOVAs were again conducted on the fusion and combination data for the /a/ and /i/ vowel

Table 4
Mean Percent Response for the Female Face and Male Face Paired With the New Female and Male Voices in the /a/ Vowel Context

Voice	Auditory Consonant	Visual Consonant	Response Category /b_/	/d_/	/th_/	/g_/	/bg_/
			Female Face				
Female	/b_/	/b_/	98	0	2	0	0
	/b_/	/g_/	11	13*	63*	12	0
	/g_/	/b_/	1	0	0	27	72*
	/g_/	/g_/	0	1	0	98	1
Male	/b_/	/b_/	99	0	1	0	0
	/b_/	/g_/	17	13*	71*	0	0
	/g_/	/b_/	1	1	0	23	75*
	/g_/	/g_/	0	0	0	100	1
			Male Face				
Male	/b_/	/b_/	97	1	1	0	1
	/b_/	/g_/	27	22*	49*	2	0
	/g_/	/b_/	0	0	0	37	63*
	/g_/	/g_/	0	2	0	93	5
Female	/b_/	/b_/	93	3	3	1	0
	/b_/	/g_/	14	42*	41*	2	1
	/g_/	/b_/	2	0	0	50	48*
	/g_/	/g_/	0	5	7	87	6

*Typical fusion or combination situation.

Table 5
Mean Percent Response for the Female Face and Male Face Paired With the New Female and Male Voices in the /i/ Vowel Context

Voice	Auditory Consonant	Visual Consonant	Response Category /b_/	/d_/	/th_/	/g_/	/bg_/
			Female Face				
Female	/b_/	/b_/	100	0	0	0	0
	/b_/	/g_/	12	46*	42*	0	0
	/g_/	/b_/	0	0	0	18	82*
	/g_/	/g_/	0	0	0	100	0
Male	/b_/	/b_/	99	0	1	0	0
	/b_/	/g_/	8	56*	35*	1	0
	/g_/	/b_/	0	0	0	22	78*
	/g_/	/g_/	0	2	0	98	0
			Male Face				
Male	/b_/	/b_/	99	0	1	0	0
	/b_/	/g_/	27	38*	35*	0	0
	/g_/	/b_/	0	1	0	41	58*
	/g_/	/g_/	0	1	0	89	10
Female	/b_/	/b_/	99	0	1	0	0
	/b_/	/g_/	24	48*	28*	0	0
	/g_/	/b_/	0	0	0	43	57*
	/g_/	/g_/	1	0	1	84	14

*Typical fusion or combination situation.

tokens. The first ANOVA, which was run on the percent "ba" responses for the fusion stimuli, indicated no significant effect of face [$F(1,22) = .298, p > .59$], a marginally significant effect for voice [$F(1,22) = 4.27$, $p < .06$], and no significant interaction of face and voice [$F(1,22) = .65, p > .42$]. The second ANOVA, which was run on the percent "ga" responses to the combination stimuli, indicated no significant effect of face [$F(1,22) = 2.12, p > .15$] or of voice [$F(1,22) = 2.75, p > .11$], nor was there a significant interaction of the two effects [$F(1,22) = .68$, $p > .41$]. The third and fourth ANOVAs, which were run on the percent "bi" responses for the fusion stimuli and the percent "gi" responses for the combination stimuli, revealed a similar pattern of results. There were no significant effects of face or voice, nor were there any significant interactions of the two effects: face [$F(1,22) = 1.59, p > .22$], voice [$F(1,22) = .04, p > .84$], and face × voice [$F(1,22) = .65, p > .42$]; and face [$F(1,22) = 3.28, p > .08$], voice [$F(1,22) = .005, p > .94$], and face × voice [$F(1,22) = .24$, $p > .62$], for the fusion and combination stimuli, respectively. Thus, the results of Experiment 2, which was conducted with two new talkers, replicated the results of Experiment 1 in demonstrating that the McGurk effect is not influenced by incongruency between the A and V signals with respect to the gender of the talker.[5]

To summarize, the results from Experiments 1 and 2 demonstrate that the magnitude of the McGurk effect is not influenced by reductions in the congruency between the A and V signals. Moreover, this finding is not dependent on the particular talker, or on the particular vowel context. A substantial McGurk effect was obtained for both faces and all four voices used in the experiments.

Finally, there was no difference in the overall magnitude of the McGurk effect across vowel context. The percentages of "b" responses for the fusion stimuli were approximately the same for both the /a/ and /i/ contexts (15.5% and 17.1%, respectively, averaged across the four voices dubbed onto the two faces). However, an examination of the pattern of responses from Experiments 1 and 2 indicates an influence of vowel context on the McGurk effect. The percentages of "d" and "th" responses, averaged across the four voices dubbed onto the two faces, were 18.1% and 58.25% for the /a/ vowel tokens, and 50.5% and 31.9% for the /i/ vowel tokens. A two-way ANOVA of these data with vowel and response as the main effects revealed a significant interaction [$F(1,7) = 21.4, p < .005$], indicating that the pattern of "d" and "th" responses was reliably different across the two vowel contexts. Thus, the results of both Experiments 1 and 2 indicate a difference in the response pattern of the McGurk effect ("d" versus "th") for the /i/ and the /a/ vowel contexts. This finding adds to the list of differences that have already been observed for the two vowel contexts (Green et al., 1988).

EXPERIMENT 3

Experiments 1 and 2 established that a discrepancy in the gender of the talkers who produced the auditory and visual tokens did not reduce the extent to which the two streams of information—acoustic and optic—were integrated in speech perception. These findings support the view that at the time at which information from the two streams mixes, information about the specific talker has been extracted from the signal and used to neutralize the phonetic information with respect to the differences produced by different talkers. It is interesting, therefore, to assess the degree to which subjects were aware of the

gender discrepancy between the auditory and visual information. Did the subjects fail to detect the discrepancy, or did they ignore the differences between the talker's face and the talker's voice? To address this issue, subjects' awareness of the incongruity between the A and V signals was examined in Experiment 3.

Method

Subjects

The subjects were 10 new undergraduate students who were either paid or given course credit as an incentive to participate. No one reported any history of a speech or hearing disorder and all had normal or corrected-to-normal vision. All were native English speakers.

Materials

The AV stimuli consisted of two new test videotapes. The first contained only the A/ba/-V/ba/ and A/ga/-V/ga/ tokens, both congruent and incongruent with respect to the gender of the talker, taken from the previous two experiments. This resulted in a total of 16 different AV stimuli (the two utterances /ba/ and /ga/ spoken by four talkers dubbed onto two different faces). The test videotape contained four repetitions of each of the 16 different stimuli for a total of 64 test trials, which were randomized. In addition, a block of 16 trials consisting of a single repetition of each stimulus was created at the beginning of the tape as a practice set. The second test videotape was created in a similar manner using the A/bi/-V/bi/ and A/gi/-V/gi/ tokens.

Procedures

Each subject was presented with both the /a/ and the /i/ tapes in a single session lasting about 20 min. The order of presentation of the tapes was counterbalanced across the 10 subjects. Each subject was instructed to judge the degree to which the voice matched the face along a 10-point scale, on which a 1 indicated that the two signals did not match at all and a 10 indicated that they were matched perfectly. The subjects rated only the degree of match between the A and V stimuli, so that they would not have to concentrate on carrying out two tasks (an identification as well as the rating) at once. The subjects were presented with the practice trials prior to the presentation of the test stimuli. They responded verbally, enabling them to maintain their attention and vision focused on the video monitor. The subjects were tested in the same room, with the equipment from the previous two experiments.

Results

Preliminary analysis indicated no difference in the mean ratings between the A/b/-V/b/ and the A/g/-V/g/ tokens, so these data were pooled for further analysis. Figure 1 presents the results for the /a/ tokens. As can be seen in the figure, the stimuli in which the female voices are dubbed onto the female face and in which the male voices are dubbed onto the male face—our FC and MC conditions—received relatively high compatibility ratings. However, the tokens in which the voices were cross-dubbed onto opposite-gender faces—our FI and MI conditions—received relatively low compatibility ratings.

A three-way ANOVA with face (female vs. male), voice (female vs. male), and talker (A vs. B) as the main effects was conducted. There was a significant effect for face [$F(1,19) = 92.7, p < .0001$]. For these stimuli, the female face received higher overall mean ratings (5.16 vs. 4.12), although why this should be the case is currently unknown. The other two main effects for voice or talker were not significant (both Fs $< .7$, $p > .4$). Of primary importance is the face × voice interaction, which was highly significant [$F(1,19) = 112.6$, $p < .0001$]. Post hoc analyses indicated that the cross-gender tokens were given reliably lower mean ratings of compatibility than were the gender-appropriate tokens ($p < .01$).[6] The gender-compatible conditions (FC and MC) did not differ from one another. Thus, subjects could perceive the incompatibility in the cross-gender stimuli. Finally, there was also a significant face × talker interaction [$F(1,19)$

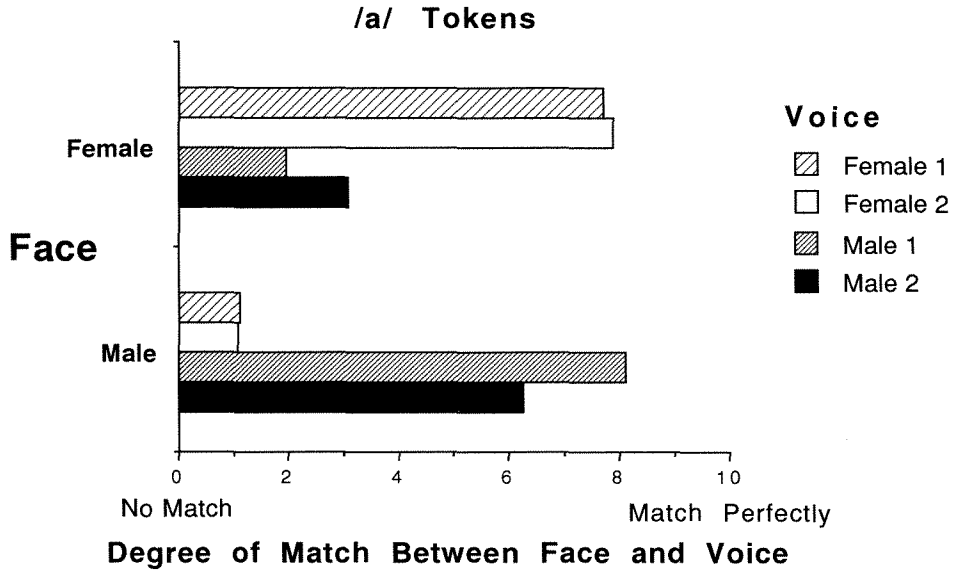

Figure 1. The mean ratings for the degree of match between the four voices and the two faces for the /a/ tokens in Experiment 3.

$= 23.3, p < .0001$], as well as a face × voice × talker interaction [$F(1,19) = 16.9, p < .0006$]. Post hoc analyses indicated that these interactions were due to the second male voice, which received significantly lower mean ratings ($p < .05$) than did the first male voice when dubbed onto the male face (6.2 vs. 8.1, respectively), and higher mean ratings ($p < .05$) when dubbed onto the female face (3.1 vs. 2.0). However, this pattern of responses did not occur for the two female voices, which did not differ reliably regardless of the face onto which they were dubbed.

A similar pattern of results occurred for the /i/ tokens (see Figure 2), and a three-way ANOVA produced similar results. There was a significant main effect for face [$F(1,19) = 14.4, p < .005$] as well as talker [$F(1,19) = 6.3, p < .05$]. Post hoc analyses indicated that the face effect was again due to the female face's receiving overall higher mean ratings than the male face did (5.3 vs. 4.6, respectively). The talker effect was due to the first set of talkers' receiving overall higher mean ratings than the second set did (5.2 vs. 4.7). The effect of voice was not significant [$F(1,19) = 1.6, p > .2$]. The critical finding is that for the /i/ tokens, just as for the /a/ tokens, the face × voice interaction was highly significant [$F(1,19) = 51.1, p < .0001$]. Post hoc analyses indicated that the mean compatibility ratings for the cross-gender tokens were again reliably lower than the gender-appropriate tokens ($p < .01$), which were not reliably different. Again, there was a significant face × talker interaction [$F(1,19) = 37.1, p < .0001$], as well as a significant face × voice × talker interaction [$F(1,19) = 34.5, p < .0001$]. Post hoc analyses indicated that the interactions with talker were again due to the second male voice, which produced reliably lower mean ratings than did the first male voice when dubbed onto the male face ($p < .05$), and higher mean ratings when dubbed onto the female face ($p < .05$). The female voices, however, did not differ significantly.[7]

The results of Experiment 3 indicate that the discrepancy between the cross-gender stimuli was quite apparent to the subjects. The subjects were very consistent in rating the normal AV stimuli as matching and the cross-gender stimuli as not matching. This strongly suggests that the results from Experiments 1 and 2 were not due to the subjects' *inability* to detect the incompatibility between the A and V signals of the cross-gender stimuli. The subjects did perceive the gender incompatibility of the signals.

GENERAL DISCUSSION

In this series of experiments, a reduction in the cognitive congruency between the auditory and the visual signals was created by dubbing a male talker's voice onto a female talker's face, and vice versa. The results of this study revealed no impact on the magnitude of the McGurk effect, even though the incongruency was quite apparent. Thus, the mechanism responsible for integrating the phonetic information is not disrupted by the cognitive discrepancies between the two signals. These results indicate that the McGurk effect does not operate in a manner similar to other types of intersensory phenomena, such as the ventriloquism effect, which is sensitive to the cognitive congruency between the auditory and the visual signals (see, e.g., Jack & Thurlow, 1973; Jackson, 1953; Warren, 1979; Warren et al., 1981). In those studies, a

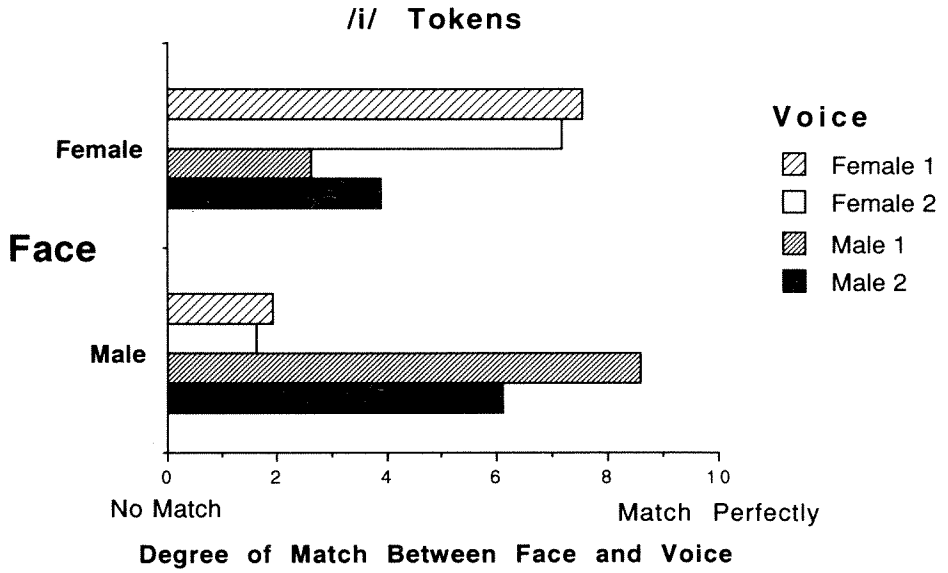

Figure 2. The mean ratings for the degree of match between the four voices and the two faces for the /i/ tokens in Experiment 3.

reduction in the cognitive congruency between the auditory and the visual signals produced a corresponding decrease in the magnitude of auditory-visual integration.

Welch and Warren (1980) proposed a general model for processing intersensory discrepancies. They suggested that the "unity" of the signals was fundamental, and they based this assumption, in part, on the results of the auditory-visual ventriloquism effect and other cross-modal phenomena. The basic notion of the unity assumption is that a greater cross-modal effect will occur in situations in which the observer assumes a single distal object or event. Welch and Warren point out a number of factors that can influence the unity assumption, including the instructions given to the subject, as well as the cognitive compellingness of the stimulus situation.

The results from the present study have important implications for Welch and Warren's (1980) model. For example, the results indicate that the unity assumption of the model may not be a precondition for the integration of all types of auditory and visual signals. Welch and Warren cite several examples of various cross-modal studies in support of their assumption, but they also suggest that certain cross-modal combinations (visual-proprioceptive) behave differently from the ventriloquism effect. The data from the present experiments show that the auditory-visual McGurk effect is resistant to the reduction in unity created in the present study. Other ways of reducing the congruency between the auditory and visual signals, such as altering the temporal or spatial synchrony, might produce lower McGurk effects as predicted by the unity assumption. However, the present data do not provide support for the generality of Welch and Warren's model.

The results of the present experiments do provide support for certain models of speech perception, in particular those concerned with how the perceptual system handles the variation in the acoustic realizations of phonetic segments spoken by different talkers. Specifically, it has been suggested on the basis of experiments on adults that the perceptual system normalizes the acoustic information with respect to talker differences at an early stage in processing and that this neutralized information is used during phonetic processing (Mullennix & Pisoni, 1990; Mullennix et al, 1989; Nusbaum, 1990; Pisoni, 1990; Syrdal & Gopal, 1986). Furthermore, developmental work indicates that, well before language production, young infants demonstrate the ability to perceptually normalize speech signals. It has been shown that 6-month-old infants can detect the constant phonetic identity of a speech signal despite changes in the gender of the talker (Kuhl, 1979, 1983). More recent data suggest that 2-month-old infants do so as well (Jusczyk, Bertoncini, Bijeljac-Babic, Kennedy, & Mehler, 1990; Marean, Werner, & Kuhl, in press). Normalization with respect to gender of the talker appears to be a very basic process in speech perception (Kuhl, 1985).

The results of the present study are compatible with this notion because they demonstrate that differences in the gender of the talker producing the auditory and visual signals had no impact on the integration of the phonetic information. Thus, by the time the phonetic information was integrated from the auditory and the visual modalities, it was sufficiently abstract as to be neutral with respect to the talker differences. Nonetheless, the results of Experiment 3 show that observers are very aware of an incompatibility between the cross-gender face-voice pairs. This suggests that the neutralization of talker differences for the purposes of phonetic categorization does not result in a loss of detailed information about the talker (cf. Pisoni, 1990).

It is interesting to compare the results of the present study with the results of studies of the dichotic presentation of speech sounds. For example, Cutting (1976) examined "psychoacoustic fusion," which bears a striking resemblance to the McGurk effect. He observed that in situations in which an auditory /ba/ is presented to one ear and an auditory /ga/ to the other ear, listeners report hearing the syllable "da" about 50% of the time. Cutting demonstrated that the number of "da" fusion responses was unaffected by differences in the fundamental frequencies between the /ba/ and /ga/ syllables.

A similar situation occurs in a phenomenon called "duplex perception," in which part of the acoustic information for a CV syllable, such as a rising or falling third formant transition, is presented to one ear while the remainder of the syllable is simultaneously presented to the other. In this situation, listeners typically report hearing two distinct percepts: a normal speech syllable corresponding to the fused speech information (e.g., /da/ if the third formant is falling and /ga/ if it is rising), and a nonspeech chirp that is either rising or falling in pitch, depending on the direction of the transition (Bentin & Mann, 1990; Cutting, 1976; Liberman, Isenberg, & Rakerd, 1981; Mann & Liberman, 1983; Nygaard & Eimas, 1990; Rand, 1974; Whalen & Liberman, 1987). Nygaard and Eimas (1990; see also Cutting, 1976) have shown that the integration of the information from the two ears is unaffected by differences in fundamental frequency of up to 256 Hz. Nygaard and Eimas (1990) conclude that since duplex perception does not require a near match in the physical properties of the stimuli, the perceptual system responsible for the integration must be working on somewhat abstract information.

Thus, for both psychoacoustic fusion and duplex perception, the integration of phonetic information presented dichotically occurs even though the acoustic information corresponds to two different talkers. As in the McGurk effect, the phonetic information that is integrated is abstracted away from the particular talker characteristics underlying the two signals. The parallels among these three phenomena again support the idea that talker normalization occurs early in speech processing. In addition, the phenomenon of duplex perception has often been considered as evidence supporting the notion of a specialized module dedicated solely to the extraction of the phonetic aspects of the speech signal (Liberman & Mattingly, 1985, 1989; Whalen & Liberman, 1987). The results of the present study are consistent with this notion, inasmuch

as they indicate that the mechanism responsible for integrating phonetic information from the auditory and visual modalities, like the mechanism responsible for duplex perception, is also unaffected by the incongruency between the signals with respect to talker characteristics. The fact that other phenomena such as the ventriloquism effect behave differently with regard to the amount of congruence between the auditory and visual signals might be due to their being handled by different mechanisms organized for the purpose, such as scene analysis or sound localization.

We return now to the central finding reported here and its implications for a broader theory of speech perception. The McGurk effect poses a challenge to all current models of speech perception. Information from two separate modalities, auditory and visual, is integrated to yield a unified speech percept. One challenge for such models is to characterize the way that the information from the two modalities is integrated. Theories of speech perception have taken two basic approaches. The first is to propose that a common metric or code unites the information from the two modalities. For example, gestural accounts (Fowler, 1986; Fowler & Rosenblum, 1991; Liberman & Mattingly, 1985, 1989) assert that during speech perception, listeners derive the articulatory movements that underlie the production of the phonetic segments. According to this account, information from both the auditory and the visual modalities is converted into a gestural code. This code then serves as the common metric for integrating and mapping the information onto underlying phonetic representations. An alternative possibility is to convert the information from the two modalities into a code that is auditorially, rather than gesturally, based. For example, Summerfield (1987) has proposed that estimates of the vocal tract filter function might be extracted from both the auditory and the visual modalities and used as a common metric for integrating the information. Other codes are also possible (cf. Summerfield, 1987), including ones that are neither strictly auditory nor strictly gesturally based but instead are amodal in nature, as Kuhl and Meltzoff (1984, 1988) and Studdert-Kennedy (1986) have advocated.

A second approach taken by some theorists is to map different metrics derived separately from the auditory and visual modalities onto underlying phonetic representations or prototypes. Several experiments have demonstrated the use of prototypes in phonetic categorization (Kuhl, 1991; J. L. Miller, Connine, Schermer, & Kluender, 1983; Samuel, 1982). An example of a prototype approach that has been applied to auditory-visual phenomena is Massaro's (1987) fuzzy logic model of perception. According to this model, information from a variety of sources is mapped onto learned phonetic prototypes.

Regardless of which of these or other approaches may eventually prove to be most useful, there are three issues that must be accommodated by all models of auditory-visual speech perception. First, as shown in these experiments and others, the McGurk effect represents a situation in which auditory and visual information that has not been previously experienced or linked together is seamlessly combined during phonetic perception. Observers do not typically experience auditory /ba/s dubbed onto visual /ga/s, yet this information is integrated during phonetic perception without any awareness of the conflict. Second, the new data reported here demonstrate that auditory and visual phonetic information is integrated even when the two inputs could not have been derived from a single, biological source. As shown in Experiment 3, subjects were aware of the fact that there was not a single event that was the source of the information they received. They readily recognized that the auditory and visual information came from two different sources, yet the processing system integrated the phonetic information anyway.

Finally, studies indicate that by the age of 4 months, infants have a rudimentary ability to relate auditorially presented speech signals and their concomitant visual gestures. Given a choice, infants will systematically look at the face that matches the speech sound they are presented with auditorially, as opposed to looking at the face that does not match the sound (Kuhl & Meltzoff, 1982, 1984; MacKain, Studdert-Kennedy, Spieker, & Stern, 1983). In short, the challenge for theories of auditory-visual speech perception is to provide adequate accounts for why auditory-visual information that observers have never experienced and that cannot stem from real biological sources are combined during phonetic perception, and moreover, to recognize that the developmental roots of the ability to relate auditory and visual speech information are in place at a very early age.

In summary, the results of the present study indicate that the McGurk effect, unlike other auditory-visual phenomena, is not influenced by reductions in the cognitive congruency between the auditory and the visual signals, even when the incompatibility is readily apparent. The reduction in congruity, created by an obvious discrepancy with respect to the talker characteristics, had no impact on the integration of phonetic information from the two modalities. This supports the notion that the differences in phonetic information that are attributable to talker variability are recoded to produce more abstract information early in the processing of speech, prior to the time that information from the auditory and visual modalities is integrated during phonetic perception. The present findings are therefore relevant to models of speech perception, as well as more general models of intermodal perception.

REFERENCES

BENGUEREL, A., & PICHORA-FULLER, M. K. (1982). Coarticulation effects in lipreading. *Journal of Speech & Hearing Research*, **25** 600-607.

BENTIN, S., & MANN, V. (1990). Masking and stimulus intensity effects on duplex perception: A confirmation of the dissociation between speech and nonspeech modes. *Journal of the Acoustical Society of America*, **88**, 64-74.

BINNIE, C. A., MONTGOMERY, A. A., & JACKSON, P. L. (1974). Auditory and visual contributions to the perception of selected English consonants for normally hearing and hearing-impaired listeners. In H. Bir

Nielsen & E. Kampp (Eds.), *Visual and audio-visual perception of speech* (Scandinavian Audiology Supplementum 4, pp. 181-209). Stockholm: Almquist & Wiksell.

COHEN, M. M. (1984). *Processing of visual and auditory information in speech perception*. Unpublished doctoral dissertation, University of California, Santa Cruz.

CUTTING, J. E. (1976). Auditory and linguistic processes in speech perception: Inferences from six fusions in dichotic listening. *Psychological Review*, **83**, 114-140.

DIXON, N. F., & SPITZ, L. (1980). The detection of auditory visual desynchrony. *Perception*, **9**, 719-721.

DODD, B. (1977). The role of vision in the perception of speech. *Perception*, **6**, 31-40.

DODD, B. (1979). Lip reading in infants: Attention to speech presented in- and out-of-synchrony. *Cognitive Psychology*, **11**, 478-484.

DORMAN, M. F., STUDDERT-KENNEDY, M., & RAPHAEL, L. J. (1977). Stop-consonant recognition: Release bursts and formant transitions as functionally equivalent, context-dependent cues. *Perception & Psychophysics*, **22**, 109-122.

ERBER, N. P. (1971). Effects of distance on the visual reception of speech. *Journal of Speech & Hearing Research*, **14**, 848-857.

FISCHER-JORGENSEN, E. (1954). Acoustic analysis of stop consonants. *Miscellanea Phonetica*, **2**, 42-59.

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., & ROSENBLUM, L. D. (1991). The perception of phonetic gestures. In I. G. Mattingly & M. Studdert-Kennedy (Eds.), *Modularity and the motor theory of speech perception* (pp. 33-59). Hillsdale, NJ: Erlbaum.

GREEN, K. P., & KUHL, P. K. (1989). The role of visual information in the processing of place and manner features in speech perception. *Perception & Psychophysics*, **45**, 34-42.

GREEN, K. P., & KUHL, P. K. (1991). Integral processing of visual place and auditory voicing information during phonetic perception. *Journal of Experimental Psychology: Human Perception & Performance*, **17**, 278-288.

GREEN, K. P., KUHL, P. K., & MELTZOFF, A. N. (1988). Factors affecting the integration of auditory and visual information in speech: The effect of vowel environment. *Journal of the Acoustical Society of America*, **84**, S155.

GREEN, K. P., & MILLER, J. L. (1985). On the role of visual rate information in phonetic perception. *Perception & Psychophysics*, **38**, 269-276.

JACK, C. E., & THURLOW, W. R. (1973). Effects of degree of visual association and angle of displacement on the "ventriloquism" effect. *Perceptual & Motor Skills*, **37**, 967-979.

JACKSON, C. V. (1953). Visual factors in auditory localization. *Quarterly Journal of Experimental Psychology*, **5**, 52-65.

JUSCZYK, P. W., BERTONCINI, J., BIJELJAC-BABIC, R., KENNEDY, L. J., & MEHLER, J. (1990). The role of attention in speech perception by young infants. *Cognitive Development*, **5**, 265-286.

KIRK, R. E. (1968). *Experimental design procedures for the behavioral sciences*. Belmont, CA: Wadsworth.

KUHL, P. K. (1979). Speech perception in early infancy: Perceptual constancy for spectrally dissimilar vowel categories. *Journal of the Acoustical Society of America*, **66**, 1668-1679.

KUHL, P. K. (1980). Perceptual constancy for speech-sound categories in early infancy. In G. H. Yeni-Komshian, J. F. Kavanaugh, & C. A. Ferguson (Eds.), *Child phonology: Vol. 2. Perception* (pp. 41-66). New York: Academic Press.

KUHL, P. K. (1983). Perception of auditory equivalence classes for speech in early infancy. *Infant Behavior & Development*, **6**, 263-285.

KUHL, P. K. (1985). Categorization of speech by infants. In J. Mehler & R. Fox (Eds.), *Neonate cognition: Beyond the blooming, buzzing confusion* (pp. 231-262). Hillsdale, NJ: Erlbaum.

KUHL, P. K. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics*, **50**, 93-107.

KUHL, P. K., & MELTZOFF, A. N. (1982). The bimodal perception of speech in infancy. *Science*, **218**, 1138-1141.

KUHL, P. K., & MELTZOFF, A. N. (1984). The intermodal representation of speech in infants. *Infant Behavior & Development*, **7**, 361-381.

KUHL, P. K., & MELTZOFF, A. N. (1988). Speech as an intermodal object of perception. In A. Yonad (Ed.), *Perceptual development in infancy: The Minnesota symposia on child psychology* (pp. 235-266). Hillsdale, NJ: Erlbaum.

LIBERMAN, A. M., ISENBERG, D., & RAKERD, B. (1981). Duplex perception of cues for stop consonants: Evidence for a phonetic mode. *Perception & Psychophysics*, **30**, 133-143.

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

LIBERMAN, A. M., & MATTINGLY, I. G. (1989). A specialization for speech perception. *Science*, **243**, 489-494.

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

MACKAIN, K., STUDDERT-KENNEDY, M., SPIEKER, S., & STERN, D. (1983). Infant intermodal speech perception is a left-hemisphere function. *Science*, **219**, 1347-1349.

MANN, V. A., & LIBERMAN, A. M. (1983). Some differences between phonetic and auditory modes of perception. *Cognition*, **14**, 211-235.

MANUEL, S. Y., REPP, B. H., STUDDERT-KENNEDY, M., & LIBERMAN, A. M. (1983). Exploring the "McGurk effect." *Journal of the Acoustical Society of America*, **74**, S66. (Abstract)

MAREAN, G. C., WERNER, L. W., & KUHL, P. K. (in press). Vowel categorization in very young infants. *Developmental Psychology*.

MASSARO, D. (1987). Speech perception by ear and eye. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip-reading* (pp. 53-83). London: 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.

MCGRATH, M., & SUMMERFIELD, 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. (1976). Hearing lips and seeing voices. *Nature*, **264**, 746-748.

MILLER, J. D. (1989). Auditory-perceptual interpretation of the vowel. *Journal of the Acoustical Society of America*, **85**, 2114-2134.

MILLER, J. L., CONNINE, C. M., SCHERMER, T. M., & KLUENDER, K. R. (1983). A possible auditory basis for internal structure of phonetic categories. *Journal of the Acoustical Society of America*, **73**, 2124-2133.

MILLS, A. E., & THIEM, R. (1980). Auditory-visual fusions and illusions in speech perception. *Linguistische Berichte*, **68**, 85-109.

MULLENNIX, J. W., & PISONI, D. B. (1990). Stimulus variability and processing dependencies in speech perception. *Perception & Psychophysics*, **47**, 379-390.

MULLENNIX, J. W., PISONI, D. B., & MARTIN, C. S. (1989). Some effects of talker variability on spoken word recognition. *Journal of the Acoustical Society of America*, **85**, 365-378.

NEAREY, T. M. (1989). Static, dynamic, and relational properties in vowel perception. *Journal of the Acoustical Society of America*, **85**, 2088-2113.

NUSBAUM, H. C. (1990). The role of learning and attention in speech perception. In H. Fujisaki (Ed.), *Proceedings of the International Conference on Spoken Language Processing* (pp. 409-412). Tokyo: Acoustical Society of Japan.

NYGAARD, L. C., & EIMAS, P. D. (1990). A new version of duplex perception: Evidence for phonetic and nonphonetic fusion. *Journal of the Acoustical Society of America*, **88**, 75-86.

PISONI, D. B. (1990). Effects of talker variability on speech perception: Implications for current research and theory. In H. Fujisaki (Ed.), *Proceedings of the International Conference on Spoken Language Processing* (pp. 1399-1407). Tokyo: Acoustical Society of Japan.

RAND, T. C. (1974). Dichotic release from masking for speech. *Journal of the Acoustical Society of America*, **55**, 678-680.

REISBERG, D., MCLEAN, J., & GOLDFIELD, A. (1987). Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip-reading* (pp. 97-113). London: Erlbaum.

REPP, B. H., & LINN, H. (1989). Acoustic properties and perception of stop consonant release transients. *Journal of the Acoustical Society of America*, **85**, 379-396.

ROBERTS, M., & SUMMERFIELD, Q. (1981). Audiovisual presentation demonstrates that selective adaptation in speech perception is purely auditory. *Perception & Psychophysics*, **30**, 309-314.

SAMUEL, A. G. (1982). Phonetic prototypes. *Perception & Psychophysics*, **31**, 307-314.

STRANGE, W. (1989). Evolving theories of vowel perception. *Journal of the Acoustical Society of America*, **85**, 2081-2087.

STUDDERT-KENNEDY, M. (1986). Development of the speech perceptuomotor system. In B. Lindblom and R. Zetterstrom (Eds.), *Precursors of early speech* (pp. 205-217). New York: Stockton Press.

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

SUMMERFIELD, 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). London: Erlbaum.

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

SYRDAL, A. K., & GOPAL, H. S. (1986). A perceptual model of vowel recognition based on the auditory representation of American English vowels. *Journal of the Acoustical Society of America*, **79**, 1086-1100.

WARREN, D. H. (1979). Spatial localization under conflict conditions: Is there a single explanation? *Perception*, **8**, 323-337.

WARREN, D. H., WELCH, R. B., & MCCARTHY, T. J. (1981). The role of visual-auditory "compellingness" in the ventriloquism effect: Implications for transitivity among the spatial senses. *Perception & Psychophysics*, **30**, 557-564.

WELCH, R. B. (1989). A comparison of speech perception and spatial localization. *Behavioral & Brain Sciences*, **12**, 776-777.

WELCH, R. B., & WARREN, D. H. (1980). Immediate perceptual response to intersensory discrepancy. *Psychological Bulletin*, **88**, 638-667.

WHALEN, D. H., & LIBERMAN, A. M. (1987). Speech perception takes precedence over nonspeech perception. *Science*, **237**, 169-171.

WOODWARD, M. F., & BARBER, C. G. (1960). Phoneme perception of lipreading. *Journal of Speech & Hearing Research*, **3**, 212-222.

NOTES

1. McGrath and Summerfield (1985) provide evidence that under the proper conditions, observers are capable of discriminating temporal asynchronies between auditory and visual speech signals of only 80 msec. However, it is unclear whether the thresholds obtained in their discrimination experiment would generalize to a situation like the McGurk effect, in which the subjects' attention is focused on the task of identifying the speech utterances, rather than on the temporal asynchrony of the auditory and visual information.

2. Some studies of the McGurk effect have employed synthetic speech dubbed onto normal faces, and it is possible that subjects in these experiments might have detected a lack of correspondence between the auditory and visual information (cf. Massaro & Cohen, 1983). Although these studies have replicated the typical McGurk effect, there have been no corresponding control conditions to determine whether the reduction in correspondence between the synthetic auditory speech and the normal talker's face had any impact on the magnitude of the effect.

3. All data analyses of the identification data were calculated using arcsine transformations of the raw data.

4. Analyses were also conducted on the fusion stimuli for both vowels, using the percent fusion responses (number of "d" and "th" responses) rather than percent "b." The results of these analyses were very similar. The important result is that again there were no significant face × voice interactions [both $Fs(1,22) < 3.0$, $p > .09$].

5. Similar analyses were run on the fusion stimuli for both vowels, using the percent fusion responses (number of "d" and "th" responses) rather than percent "b." These analyses also indicated no significant face × voice interactions [both $Fs(1,22) < 1.51$, $p > .23$].

6. All post hoc analyses were done with the Neuman-Keuls test (Kirk, 1968).

7. It is not clear why the second male voice was less compatible with the male face and more compatible with the female face than the first male voice was. One possibility is that the fundamental frequency of the second male voice was higher than the first and thus more compatible with the perception of a female talker. An analysis of the fundamental frequencies of the two voices indicated that the second male voice was slightly higher than the first (114 vs. 101 Hz). Interestingly, the difference in fundamental frequency between the two female talkers, who did not differ in their compatibility ratings, was much smaller (160 and 163 for the first and second talkers, respectively). Although the second male voice was more compatible with the female face than the first male voice was, it was clearly perceived as a male voice. In a follow-up test, we presented the same stimuli that were used in the rating experiment to a new group of 5 subjects who were simply asked to say, on each trial, whether the face was male or female, and whether the voice was male or female. All of these subjects were highly accurate (greater than 98%) at correctly specifying the gender of the face and the voice for *all* the stimuli used in Experiment 3, including the second male voice. The few errors that were made were scattered evenly across the four voices. The faces were identified correctly 100% of the time.

8. An additional assumption often held by people who take this alternative approach is that speech perception is accomplished with general auditory processes that are also involved in the perception of other auditory signals. Therefore, no special mechanism or module devoted exclusively to speech is required.

APPENDIX

The following instructions were read to the subjects in Experiments 1 and 2:

> You will be participating in an experiment concerning the perception or understanding of speech sounds. During the experiment you will be watching the video monitor in front of you. We will show you a videotape of a talker saying various syllables. The talker will be saying one of the syllables listed on the sheet in front of you: either "b, d, g, th, or bg." Your task will be to indicate what you *heard* the talker say. Simply repeat aloud what you heard the talker say after every trial. An assistant will be sitting in the room with you, to record your responses.
>
> It is important that you make a response as quickly and as accurately as you can on every trial. The trials occur every 6 seconds, so it is important that you respond and then get ready for the next trial. If you are not exactly sure what you heard, then make your best guess.

(Manuscript received December 19, 1990;
revision accepted for publication July 16, 1991.)