Perception & Psychophysics
1996, 58 (3), 351–362

Temporal constraints on the McGurk effect

K. G. MUNHALL
Queen's University and ATR Human Information Processing Research Laboratories
Kingston, Ontario, Canada

P. GRIBBLE
McGill University, Montreal, Quebec, Canada

and

L. SACCO and M. WARD
Queen's University, Kingston, Ontario, Canada

Three experiments are reported on the influence of different timing relations on the McGurk effect. In the first experiment, it is shown that strict temporal synchrony between auditory and visual speech stimuli is not required for the McGurk effect. Subjects were strongly influenced by the visual stimuli when the auditory stimuli lagged the visual stimuli by as much as 180 msec. In addition, a stronger McGurk effect was found when the visual and auditory vowels matched. In the second experiment, we paired auditory and visual speech stimuli produced under different speaking conditions (fast, normal, clear). The results showed that the manipulations in both the visual and auditory speaking conditions independently influenced perception. In addition, there was a small but reliable tendency for the better matched stimuli to elicit more McGurk responses than unmatched conditions. In the third experiment, we combined auditory and visual stimuli produced under different speaking conditions (fast, clear) and delayed the acoustics with respect to the visual stimuli. The subjects showed the same pattern of results as in the second experiment. Finally, the delay did not cause different patterns of results for the different audiovisual speaking style combinations. The results suggest that perceivers may be sensitive to the concordance of the time-varying aspects of speech but they do not require temporal coincidence of that information.

When the face moves during speech production it provides information about the place of articulation as well as the class of phoneme that is produced. Evidence from studies of lipreading as well as studies of speech in noise (e.g., Sumby & Pollack, 1954) suggest that perceivers can gain significant amounts of information about the speech target through the visual channel. How this information is combined with speech acoustics to form a single percept, however, is not clear. One useful approach to studying audiovisual integration in speech is to dub various auditory stimuli onto different visual speech stimuli. When a discrepancy exists between the information from the two modalities, subjects fuse the visual and auditory information to form a new percept. For example, when the face articulates /gi/ and the auditory stimulus is /bi/, many subjects report hearing /di/. This phenomenon has been called the McGurk effect (McGurk & MacDonald, 1976), and in this paper we use this effect to study audiovisual speech perception.

Since the original report on the McGurk effect (McGurk & MacDonald, 1976), there have been numerous replications of the phenomenon (e.g., Green & Kuhl, 1989, 1991; Green, Kuhl, & Meltzoff, 1988; Green, Kuhl, Meltzoff, & Stevens, 1991; MacDonald & McGurk, 1978; Manuel, Repp, Studdert-Kennedy, & Liberman, 1983; Massaro, 1987; Massaro & Cohen, 1983; Sekiyama & Tohkura, 1991; Summerfield & McGrath, 1984). These papers have reported a number of basic facts about the McGurk effect, including that it is influenced by the vowel context that consonants are spoken in (Green et al., 1988), that vowels themselves can show McGurk effects (Summerfield & McGrath, 1984), that the visual information for place of articulation can influence the auditory perception of consonants that differ in voicing (Green & Kuhl, 1989), and so on. However, as Green et al. (1991) pointed out, these papers have not described the necessary and sufficient conditions under which audiovisual integration occurs. Here we present three experiments that try to clarify some of the temporal influences on the McGurk effect.

Timing in Audiovisual Integration

It is a common experience when watching dubbed foreign language movies to quickly notice the disparity be-

This work was supported by a grant from NSERC and NIH Grant DC-00594. The authors acknowledge the helpful comments made by J. Magnuson, P. Thompson, Y. Tohkura, M. Tsuzaki, E. Vatikiotis-Bateson, D. Massaro, K. Green and Q. Summerfield. We thank L. Coady for testing subjects in Experiment 3 and J. Jones for doing the video analysis in Experiment 2. Correspondence should be addressed to K. G. Munhall, Department of Psychology, Queen's University, Kingston, Ont., Canada K7L 3N6 (e-mail: munhallk@qucdn.queensu.ca).

tween the auditory and visual events. The viewer immediately has a sense that the information from the two modalities comes from two different sources. In part, this perception is caused by gross disparities in the timing of the visual and auditory speech signals. It is clear that for brief, nonverbal stimuli, people are very sensitive to intermodal timing (see, e.g., Hirsh & Sherrick, 1961), with timing differences of less than 20 msec being detected. However, studies of the effects of desynchrony on the audiovisual perception of speech have reported a wide range of threshold values that are much larger than the values reported for simple transients such as clicks. Dixon and Spitz (1980) asked subjects to adjust the timing of the audio signal to match the visual signal for connected speech stimuli. They found that, on average, the audio lag had to be greater than 250 msec before subjects noticed the discrepancy. Similar time values were found by Koenig (1965, cited in McGrath & Summerfield, 1985) in an experiment in which visual stimuli were combined with low-pass filtered speech.

McGrath and Summerfield (1985) presented subjects with audiovisual sentences in which the audio track was replaced by a pulse train derived from an electroglottograph signal. Thus, the audio track provided information only about the prosodic features of the sentences and information about the timing of voicing onset and offset. On average, their subjects showed no decrease in accuracy of transcription of the sentences with the audio track delayed 20, 40, and 80 msec. However, there was a reliable decrease in transcription performance when the audio was delayed 160 msec.

For multidimensional stimuli such as speech, however, it may not be useful to try to establish *exact* detection limens for audiovisual synchrony without a more explicit characterization of the stimulus. In experiments of the kind described above, stimuli can differ along so many different perceptual or informational dimensions that estimates of the threshold for audiovisual desynchrony may vary considerably. However, what is clear from the existing data is that the delays required to disrupt speech perception are surprisingly large. Although a great deal of evidence from the study of the auditory perception of speech indicates that we are sensitive to small temporal differences in acoustic intervals (see Miller, 1986, for a review of timing effects in speech), the values reported by Dixon and Spitz (1980) and others for audiovisual timing are in the syllable or demisyllable range. This fact has important practical and theoretical implications (McGrath & Summerfield, 1985).

From a practical point of view, the large delay values are useful for any aural rehabilitation aid that involves significant amounts of signal processing. From a theoretical point of view, the delays raise questions about the conditions for audiovisual integration in speech and the stage at which the information combines. An integration process that occurs prior to any higher level speech processing would require some physical basis on which to combine the two information channels. For example, one possibility is that the source of the information from the two modalities would have to be matched and intermodal integration would be influenced by the extent to which the two modalities were physically correlated (e.g., same spatial location or same movement in space, same point in time or same variation in time). Welch and Warren (1980) proposed such a model, which required perceptual unity for integration of information from different modalities. Recently, Green et al. (1991) have shown that one aspect of perceptual unity, namely, knowing whether the information from two modalities corresponded, was not a precondition for perception of the McGurk effect. In the Green et al. study, subjects viewed stimuli composed of faces and voices of different genders. When male faces were combined with female voices and vice versa, subjects showed no decrease in the magnitude of the McGurk effect even though it was clear that the genders of the face and voice were incompatible. In three experiments here we explore how the temporal congruence of the visual and auditory information influences the McGurk effect.

GENERAL METHOD

Subjects

The subjects were native speakers of Canadian English. Different subjects served in each of the three experiments.

Stimulus Materials

The stimuli for all experiments consisted of visual /ɑgɑ/ or /igi/ paired with audio /ɑbɑ/. The visual stimuli were stored on videodisc. In Experiment 1 the images were from the Bernstein and Eberhardt (1986) database. In Experiments 2 and 3 the images were from a videodisc recorded at Queen’s University. The auditory stimuli were digitized from the original sound tracks of the videodiscs at a 22-kHz sampling rate using a 12-bit A/D board (DataTranslation, DT2820). In all three of the experiments we used natural productions of VCV stimuli.

Equipment

Subjects watched the displays on a 20-in. video monitor (Sony Model PVM 1910) and the videodiscs were played on a Pioneer (Model LD-V8000) videodisc player. The acoustics were amplified, filtered with a 10-kHz cutoff using Frequency Devices (Model 901F1) analog filters, and played through an MG Electronics Cabaret speaker that was placed directly below the monitor. Custom software was used to control the videodisc trials, play the auditory stimuli synchronously with the video, and record subjects’ responses from the keyboard.

Synchronization of Stimuli

During the development of each experiment, the audio and visual stimuli were synchronized using the original sound track from the visual stimuli. For a face saying /ɑgɑ/ we aligned the timing of the acoustic burst onset of the /g/ from the soundtrack of the /ɑgɑ/ video with the burst onset of the auditory stimulus /b/. This timing relation was considered synchronous and the experimental software allowed this timing relationship to be reliably reproduced. The software allowed the timing of the auditory stimuli to be set for a given trial with approximately ±1-msec accuracy across trials.

Analysis of Video Images

To estimate the kinematic information available to the subjects in the visual stimuli, we used a Peak Performance video analysis system (Scheirman & Cheetham, 1990) to measure the vertical motion of the upper and lower lip. The Peak Performance system is an interactive digitizing system that allows the coordinates of manually placed cursor positions to be written to disk. The video sequences are analyzed field by field, yielding a sampling rate of 60 Hz. Points on the vermilion border of the lips in the midline of the mouth were measured and the lower lip position was subtracted from the upper lip position. This measure provides a crude measure of the change in the oral aperture (Abry & Boë, 1986). The lips, of course, are not active articulators in /g/ production, but the lower lip passively moves with the mandible, which is involved in the production of /g/. The lips and mandible move with similar timing characteristics in speech, though they will be slightly out of phase (Gracco & Abbs, 1986). The lip aperture was chosen because it can be measured reliably, because the changing oral aperture accounts for a large proportion of the visible facial motion in speech, and because listeners in audiovisual communication fix their gaze on the mouth (Vatikiotis-Bateson, Eigsti, & Yano, 1994).

Procedure

The subjects were tested individually in a large laboratory room. Subjects were seated approximately 2 m in front of the video monitor with a keyboard placed in front of them. They were instructed to watch the faces of the talkers and to listen to the output from the audio speaker and report what the audiovisual stimuli *sounded* like. They responded by choosing one of four labeled keys. Four consecutive keys on the keyboard were labeled B, D, G, and O. The first three labels stand for the stops /b/, /d/, /g/, and the final label stands for "other." Following the presentation of instructions, the subjects were given a short practice session to familiarize them with the experimental protocol. The experiments were response paced with a new trial being presented 2 sec following the subject's response. Between trials, the screen was blackened.

EXPERIMENT 1

The question addressed in this first experiment is how the McGurk effect is influenced by the temporal alignment of the auditory and visual channels. The results from a number of studies indicate that the speech perceptual system does not require a tight timing relationship between the two modalities. Cohen (1984) and Massaro and Cohen (1993) manipulated temporal asynchrony to study how visual /bɑ/ and auditory /dɑ/ are combined to be perceived as /bdɑ/. Subjects perceived /bdɑ/ even when the auditory /dɑ/ preceded the visual /bɑ/ by as much as 200 msec. As Massaro (1987) has concluded, the time of arrival of the auditory and visual information does not seem to be the critical factor in determining the percept. Tillmann, Pompino-Marschall, and Porzig (1984) have shown that for German subjects, the combination of a visual "gier" and auditory "bier" produces the McGurk percept "dier" across a wide range of temporal alignments. Tillmann et al. (1984) varied the temporal alignment of the auditory stimuli by ±500 msec. They reported that /d/ responses exceeded /b/ responses over a wide range of values (±250 msec). More recently, Ward (1992) reported that auditory delays of up to 300 msec still produced a significant number of McGurk responses.

The present study aims to replicate the study of Tillmann et al. (1984) and Ward (1992) using a different set of asynchronies with nonsense bisyllables. In addition, we manipulated the vowel quality in the visual stimuli. By using /i/ and /ɑ/ vowel contexts, we presented the subjects with two different patterns of visual motion.

Method

Subjects. Nineteen undergraduates at Queen's University served as subjects. The subjects were native speakers of Canadian English and reported no speech, language, or hearing problems. All had normal or corrected-to-normal vision. Four subjects were eliminated because they gave the same response for all trials and added no information to the experiment. Three of these subjects answered /b/ for all stimuli and thus never perceived the McGurk effect. The 4th of these subjects responded /d/ for all stimuli and delays. Thus, data analyses were carried out on 15 subjects.

Stimuli. The visual stimuli were the female speaker's production of /igi/ and /ɑgɑ/ from the Bernstein and Eberhardt (1986) videodiscs. The auditory stimulus was a digitized version of the same speaker's productions of /ɑba/. The timing of the auditory stimuli varied in 60-msec steps from 360 msec prior to synchrony to 360 msec after synchrony. Thus, there were 13 audiovisual pairings for each of the two vowel contexts. The subjects responded to 10 blocks of stimuli with the 26 audiovisual pairings randomized within a block.

Results and Discussion

The dependent variable was the percentage of /b/ responses. This dependent measure indicates the degree to which the stimuli elicit the McGurk effect. The more /b/ responses, the weaker the McGurk effect.[1] A two-way repeated measures analysis of variance (ANOVA) (audiovisual synchrony × vowel) was used to analyze the data. The overall results are plotted in Figure 1. As can be seen, the vowel and synchrony conditions influenced the percentage of /b/ responses. There was a significant effect for delay [$F(12,168) = 8.57$, $p < .001$], with the large asynchronies producing higher rates of /b/ responses, and also a significant effect for vowel context [$F(1,14) = 5.85$, $p < .05$], with the visual vowel /i/ producing higher rates of /b/ responses.

The vowel effect was opposite to that reported by Green, Kuhl, and Meltzoff (1988). In their study, the vowel /i/ produced the greatest number of McGurk responses. In the Green et al. study, however, the visual and auditory vowels were the same. In the present data, the /i/ visual stimulus was paired with an auditory /ɑ/ stimulus. Thus, we cannot determine if the source of the difference was the relative effectiveness of the visual /i/ stimuli used in the experiments or the interaction of different auditory and visual information in the present experiment. It is known that some speakers are visually more intelligible than others (e.g., Gagne, Masterson, Munhall, Bilida, & Querengesser, 1994) and that speakers differ greatly in the pattern of this intelligibility across different words. In part, these differences are caused by the amount of movement for a given syllable—the greater the movement, the more intelligible the syllable. In Figure 2, the kinematics of the oral aperture are plotted for the /ɑgɑ/ and /igi/ visual stimuli used in the experiment with the traces lined up at acoustic burst onset. The aperture moves from an initial closed position to the peak opening for the first vowel. Then, it closes somewhat for

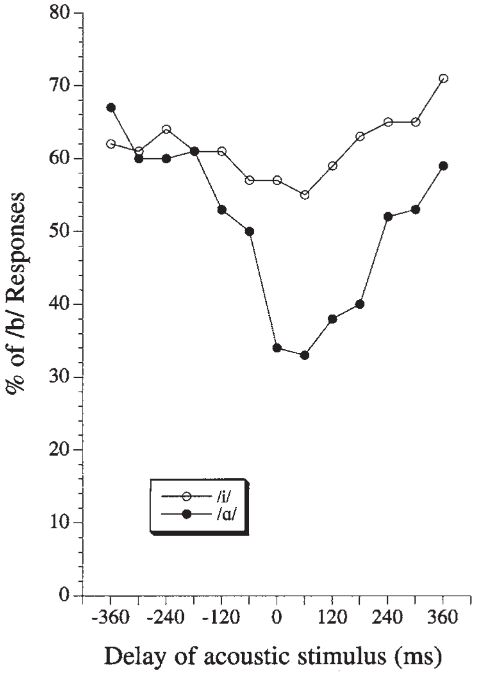

Figure 1. Experiment 1: The percentage of /b/ responses as a function of the delay of the auditory stimuli. Negative numbers on the abscissa indicate that the auditory stimulus preceded the visual stimulus. Data for the two vowel contexts are plotted separately.

the intervocalic consonant, /g/, and opens for the second vowel. Finally, the mouth closes following the end of the bisyllable. As can be seen, the amount of visual motion for the intervocalic /g/ was less in the /i/ than the /ɑ/ context in the stimuli used in this experiment.[2] This suggests that the /g/ might be visually more intelligible in the /ɑ/ context. However, Green and Gerdeman (1995) have recently shown findings similar to those reported here, suggesting that the effect may result from the vowel mismatch between auditory and visual information. When an auditory /ba/ was dubbed onto a visual /gi/, the number of "b" responses increased by over 30% compared to the matched vowel case.

The possibility that the smaller McGurk effect for the visual /igi/ in the present data is due to the mismatch of visual and auditory information is an intriguing one. It may be that the rate of change and amount of visual motion must be matched with the auditory changes to get strong audiovisual fusions. This possibility was explored directly in Experiment 2.

The synchrony manipulation produced a V-shaped function for the rate of /b/ responses. There were reliable positive linear trends [$F(1,168) = 63.34, p < .001$; $F(1,168) = 15.83, p < .001$] for both /ɑ/ and /i/, respectively, for the conditions following zero. There was a reliable negative linear trend [$F(1,168) = 19.84, p < .001$] for the conditions preceding zero only for the vowel /ɑ/. This pattern produced a significant audiovisual synchrony × vowel interaction [$F(12,168) = 5.65, p < .001$].

The function shown in Figure 1 is not symmetrical around the 0-delay axis. For the vowel /a/, there was a tendency to respond /b/ more frequently when the audio signal led the video than vice versa [$F(1,168) = 60.54$, $p < .001$]. In fact, the lowest rate of /b/ response occurred when the audio lagged the video by 60 msec rather than when the audio signal was synchronized with the sound track of the video signal. This trend is not surprising since the relative speeds of sound and light would produce many natural occurrences of auditory events lagging their visual counterparts in the natural world. For example, if someone were 30 m away from the person he/she was speaking to, the acoustics would reach the listener about 88 msec after sight of the corresponding facial movements. Summerfield (1992); Smeele, Sittig, and Van Heuven (1992); and Dixon and Spitz (1980) have reported similar asynchronies to the one we have observed.

The overall pattern of results is consistent with the data of Cohen (1984); Tillmann et al. (1984); Ward (1992); Massaro and Cohen (1993); and Green, Stevens, and Kuhl (1994). The McGurk effect does not require strict synchrony in the timing of the information from the two modalities. We used Dunnett's procedure (Dunnett, 1955) for pairwise comparisons to identify the first condition that reliably differed from the in-synchrony condition. The percentage of /b/ responses was reliably higher ($p < .05$) than the zero lag condition when the auditory stimuli were advanced by 60 msec with respect to the visual stimuli and when the auditory stimuli were delayed by 240 msec for the vowel /ɑ/. The data for the vowel /i/ were not examined because the function was so flat.

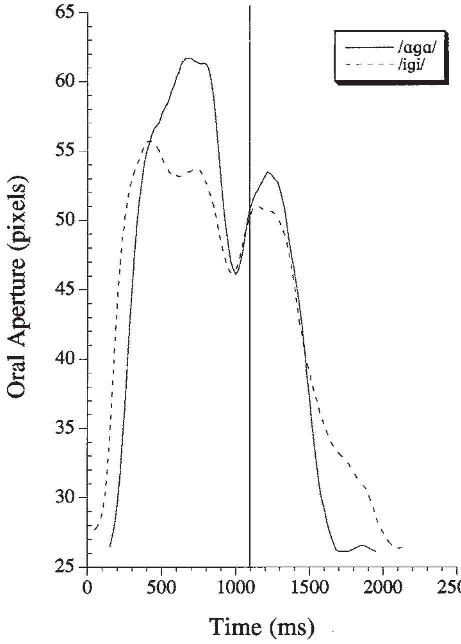

Figure 2. Kinematics of the oral aperture in the visual stimuli used in Experiment 1. Data for the two vowel contexts are plotted separately. The traces are lined up at the onset of the acoustic burst.

Table 1
Mean Percentage of /d/ and /g/ Responses as a Function of Vowel Context and Delay in Experiment 1

	/i/				/ɑ/			
	/d/		/g/		/d/		/g/	
Delay	*M*	*SEM*	*M*	*SEM*	*M*	*SEM*	*M*	*SEM*
−360	6.0	3.3	32.0	8.4	8.7	4.2	24.0	5.8
−300	4.0	3.3	34.7	7.4	8.0	3.1	32.0	6.3
−240	4.7	1.9	31.3	6.7	8.7	3.1	31.3	6.4
−180	4.7	2.4	34.7	7.8	6.7	2.9	32.3	5.4
−120	2.7	1.5	36.7	7.2	5.3	2.4	41.3	5.8
−60	2.0	2.0	41.3	8.6	7.3	3.4	42.7	6.2
0	4.7	1.9	38.0	7.9	10.7	4.0	55.3	5.9
60	3.3	2.1	41.3	8.9	6.7	3.0	60.7	6.7
120	3.3	1.9	37.3	8.0	8.7	3.2	53.3	6.5
180	2.7	1.5	34.7	7.9	8.0	3.3	5.2	7.3
240	3.3	2.1	32.0	6.6	6.0	2.7	42.0	6.3
300	1.3	0.9	34.0	8.0	4.0	2.1	43.3	7.2
360	3.3	4.9	26.0	6.5	9.3	4.2	31.3	5.6

The percentages of /d/ and /g/ responses are shown in Table 1.[3] Overall, the subjects responded /g/ at a much higher rate than /d/. In the /g/ responses there were reliable effects of delay [$F(12,240) = 8.32$, $p < .001$] and delay × vowel context interaction [$F(12,240) = 5.20$, $p < .001$]. There were no reliable patterns in the /d/ responses.

Two final aspects of the data warrant comment. As can be seen in Figure 1, the subjects never showed 100% /b/ responses even when the audio signals were 360 msec out of synchrony. Since we did not run an auditory-only condition, it is possible that the large asynchrony values represent the baseline responding level for the auditory stimuli in the experiment. This is unlikely, however, because auditory-only tests under these conditions in the same laboratory have shown very high rates of /b/ responses. Further, Tillmann et al. (1984) showed a similar response pattern in their study for the large asynchronies. It may be that the presence of simultaneous visual information, no matter what its phonetic character, can influence the auditory perception of /b/. However, it should be noted that since this is a within-subject design, it is the relative performance in the different conditions that is important.

A final issue is that the subjects in this experiment showed considerable variability in the degree to which they were subject to the McGurk effect. In the extreme, 3 of the 4 subjects who were dropped from the experiment did not experience the McGurk effect at all. The source of this variability is not known but it is not unique to the McGurk effect. Pick, Warren, and Hay (1969) reported that there seemed to be a bimodal distribution of subjects when the effects of vision on auditory location was evaluated. Some of the subjects showed a great deal of visual biasing and others showed little effect.

EXPERIMENT 2

In Experiment 1, synchrony was defined with respect to a particular moment, the onset of the release burst. This is an important point in time for stop consonant production and perception, but the information for the stop is not localized at any single point in time (see, e.g., Kashino & Craig, 1994; Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). Rather, the information for a stop extends in time throughout the preceding and following vowels. In both the visual and auditory modality, this temporally extended information derives from the moving vocal tract and is thus dynamic in nature. In Experiment 1, the possibility arose that audiovisual integration was greater when information in the two modalities was consistent. In Experiment 2, we investigated the use of this dynamic information. To this end, we manipulated speaking rate auditorily and visually and combined the different speaking rates in a factorial design. Speaking rate manipulations produce changes in the duration, velocity, and displacement in speech movements (Gay, 1981), and produce changes in the duration, slope, and extent of the formant transitions (Gay, 1978; Miller & Baer, 1983). Others have shown that the rate of visual speech information can influence the perception of auditory speech categories. Green and Miller (1985), for example, showed that the perceived boundary along a continuum of auditory voiced/voiceless stimuli could be influenced by the rate of movement of the face that was presented with the stimuli.

If matching the dynamics of the two modalities is important for successful integration, we would expect that audiovisual pairings produced at the same speaking rate would show a greater number of McGurk responses than pairings of stimuli from different speaking rates. Further, we would expect that when the speaking rates in the two modalities differ more, fewer McGurk responses will be observed. On the other hand, if the percept does not depend on the concordance of the information from the two modalities, then the degree of integration may be determined by other criteria (e.g., the relative strength or intelligibility of the information in the two modalities).

Method

Subjects. Thirty undergraduates at Queen's University served as subjects. The subjects were native speakers of Canadian English and reported no speech, language, or hearing problems. All had normal or corrected-to-normal vision.

Stimulus materials. The Queen's University videodisc contains 9 speakers who produce VCV and V utterances in three different speaking conditions. For this experiment, 3 female speakers were chosen who varied in the amount of facial motion that was used for the production of the utterances. Pilot data indicated that this variation influenced the strength of the McGurk effect. The visual stimuli were /ægæ/ utterances produced by the speakers in three different speaking conditions: fast, normal, and clear. The clear speaking condition was induced by asking the speakers to speak "more clearly," as if they were speaking to someone who was having difficulty understanding. This may have produced effects on the speech other than simply a change in rate (e.g., Lindblom, 1990; Picheny, Durlach, & Braida, 1985); however, it allowed us to produce an utterance with a longer duration naturally without resorting to the highly unnatural instruction, "speak slowly." The average acoustic durations across speakers of the /ægæ/ utterances used for visual stimuli were 304.85, 399.12, and 491.44 msec for fast, normal, and clear, respectively.

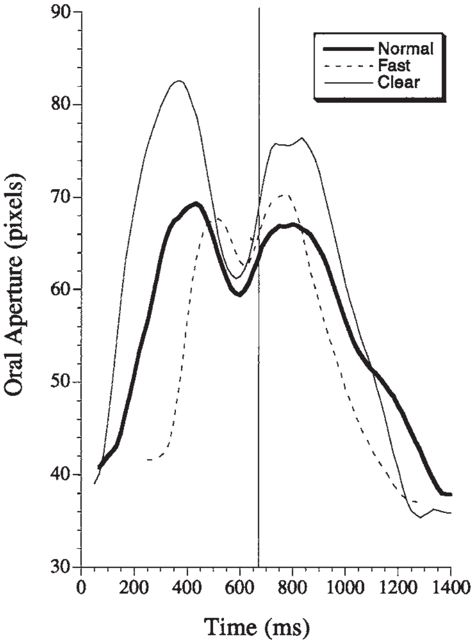

Figure 3. Kinematics of the oral aperture for the visual stimuli for speaker M.J. used in Experiment 2. Data for the three speaking conditions are plotted separately. The traces are lined up at the onset of the acoustic burst.

The auditory stimuli were productions of /æbæ/ produced by the 3 speakers in the same three speaking conditions. The average durations across speakers of the auditory /æbæ/ stimuli were 302.89, 378.18, and 488.15 msec for fast, normal, and clear, respectively. Figure 3 shows the kinematic patterns for the motion of the oral aperture for Speaker M.J. The traces plot the motion of the mouth from a closed position through the bisyllable and back to the closed mouth position. As in Figure 2, the trajectories are lined up at the onset of the acoustic release burst. As can be seen, the rate and size of the facial movements varied across speaking conditions.[4] The auditory stimuli were digitized from the sound track of the Queen's University videodisc.

During the experiment, the three visual and three audio tokens produced by each speaker were paired so that the utterance for each visual speaking condition was presented with the utterance for each auditory speaking condition. This produced nine pairings for each of the 3 speakers and thus 27 audiovisual stimuli in each block. The subjects were shown five blocks of stimuli with the 27 stimuli randomized within block. All of the auditory stimuli were timed so that the onset of the release burst in the /b/ was synchronized with the onset of the release burst in the sound track of the /g/ used as a visual stimulus.

Results and Discussion

As in Experiment 1, the main dependent variable was the percentage of /b/ responses. The data were analyzed in a three-way, repeated measures ANOVA (speaker × visual speaking condition × audio speaking condition). The 3 speakers differed in the percentage of /b/ responses that they elicited [$F(2,58) = 10.27, p < .01$]. As can be seen in Table 2, Speaker M.J.'s stimuli yielded the least /b/ responses, and Speaker L.J. produced the greatest number of /b/ responses. There were also main effects for visual speaking condition [$F(2,58) = 36.68, p < .01$] and auditory speaking condition [$F(2,58) = 17.11, p < .01$]. As the speaking rate moved from fast to normal to clear, the information within a modality increased in influence. In the auditory channel, this manifested as an increased number of /b/ responses. The auditory fast rate produced an overall average of 17.22% /b/ responses and the clear condition produced 27.96%. In the visual channel, this manifested as a decreased number of /b/ responses. The visual fast condition produced an overall average of 30.52% /b/ responses and the clear condition produced 16.26% /b/ responses (Figure 4).

If the concordance between the information from the two modalities is important for producing audiovisual integration, then we would expect that there would be a visual speaking condition × auditory speaking condition interaction. In addition, we would expect that one source of this interaction would be a lower rate of /b/ responses when audio and visual speaking conditions were matched. As predicted, there was a visual speaking condition × auditory speaking condition interaction [$F(4,116) = 7.02, p < .01$]. We examined whether the source of this interaction could be due to the concordance of the audiovisual stimuli using orthogonal contrasts. The means of the matched and unmatched audiovisual conditions differed [$F(1,116) = 15.62, p < .01$], with the three matched conditions producing, on average, fewer /b/ responses than the six unmatched conditions. In addition, the average of the two conditions that were most discordant (visual fast/auditory clear; visual clear/auditory fast) produced a higher rate of /b/ responses than the average of the four unmatched conditions that were closer to each other in rate [$F(1,116) = 45.12, p < .01$]. The mean percentages of /b/ responses for the matched (visual fast/auditory fast, visual normal/auditory normal, visual clear/auditory clear), discordant (visual fast/auditory normal, visual normal/ auditory fast, visual normal/auditory clear,

Table 2
Mean Percentage of /b/ Responses as a Function of Speaker, Visual Speaking Condition, and Auditory Speaking Condition in Experiment 2

	Visual Rate																	
	M.J.						P.B.						L.J.					
Auditory	Fast		Normal		Clear		Fast		Normal		Clear		Fast		Normal		Clear	
Rate	*M*	*SEM*	*M*	*SEM*	*M*	*SEM*	*M*	*SEM*	*M*	*SEM*	*M*	*SEM*	*M*	*SEM*	*M*	*SEM*	*M*	*SEM*
Fast	16.0	4.8	11.3	3.8	10.0	3.8	16.7	5.0	10.0	3.1	6.7	2.4	36.7	6.9	22.0	6.0	25.7	6.2
Normal	23.3	5.2	11.0	4.0	9.0	3.2	30.7	6.5	18.0	5.8	17.3	4.7	30.0	6.7	18.0	5.0	19.3	5.1
Clear	22.0	5.3	9.0	4.4	5.0	3.0	37.0	6.2	24.0	5.3	12.3	3.3	62.3	7.8	39.0	7.2	41.0	7.3

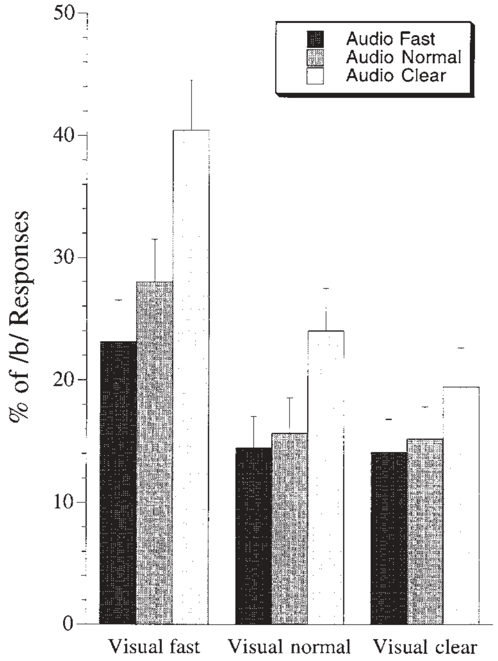

Figure 4. Experiment 2: The percentage of /b/ responses as a function of the different visual and auditory speaking condition combinations. The error bars show the standard errors of the means.

visual clear/ auditory normal), and most discordant conditions (visual fast/auditory clear, visual clear/auditory fast) were 19.41%, 20.42%, and 27.28%, respectively. Thus, the percentage of /b/ responses increased systematically as the visual and auditory information became more dissimilar.

The observed main effects of visual speaking condition and auditory speaking condition were not due to this interaction. Using contrasts orthogonal to the concordance contrasts described above (i.e., the comparison of the matched and unmatched and the comparison of the discordant and most discordant), it was found that the linear effect of auditory speaking condition [$F(1,116) = 124.40$, $p < .001$] and the linear effects of visual speaking condition [$F(1,116) = 219.25$, $p < .001$] contributed independent variance.

Although the effect of visual concordance is reliable and independent of the main effects, we note that it is not large. The two contrasts testing the concordance effect accounted for only 28% of the total variance for the visual and auditory speaking conditions and their interaction.

We did not explore interactions involving the speakers because we know little about the characteristics of individual speakers that make them more or less intelligible (Gagne et al., 1994) and because they did not seem to change any of the major patterns. The speaker × visual speaking condition × auditory speaking condition interaction was not reliable [$F(8,232) = 1.44$, $p > .1$]. The speaker × auditory speaking condition interaction was significant [$F(4,116) = 10.85$, $p < .01$]. The speaker × visual speaking condition interaction was also reliable [$F(4,116) = 2.72$, $p < .05$].

The mean percentages of /d/ and /g/ responses are shown in Table 3. In contrast to what occurred in Experiment 1, subjects responded /d/ at higher rates in this experiment than they responded /g/. It is not clear what the source of the difference between the two experiments was. Different talkers were used for the stimuli in Experiments 1 and 2 (see note 3), and thus the perceptual characteristics of the stimuli may have differed. Analyses of the /d/ and /g/ responses showed reliable speaker, visual speaking condition, and auditory speaking condition effects ($p < .01$). There was also a visual speaking condition × auditory speaking condition interaction for /d/ ($p < .01$).

In summary, the data show significant influences of visual and auditory speaking rate. For both the visual and auditory stimuli, the information within each modality influenced perception more in the clear speaking condition. In addition, there was a small but reliable tendency for the better matched stimuli to elicit more McGurk responses than unmatched conditions.

EXPERIMENT 3

In Experiment 2 we maintained synchrony at the point of acoustic release while we manipulated the dynamics of articulation. The onset and offset synchrony of the vowels, however, could not be maintained in this manipulation using natural productions. This means that the most discordant audiovisual dynamics were also the most discordant in terms of overall duration and timing of the onsets and offsets. There are two ways that this confound can be addressed. First, we could use edited or synthetic speech that is equated for acoustic duration. Second, we could directly manipulate synchrony, as in Experiment 1, for different audiovisual rates. By doing this, we could

Table 3
Mean Percentage of /d/ and /g/ Responses as a Function of Visual and Auditory Speaking Condition in Experiment 2

	Auditory Condition											
	/d/						/g/					
Visual	Fast		Normal		Clear		Fast		Normal		Clear	
Condition	*M*	*SEM*	*M*	*SEM*	*M*	*SEM*	*M*	*SEM*	*M*	*SEM*	*M*	*SEM*
Fast	36.1	4.1	40.1	3.8	38.8	4.1	37.4	4.0	25.9	3.1	14.2	2.5
Normal	39.9	4.4	44.7	4.1	47.3	4.3	42.8	4.3	35.0	3.8	24.0	3.1
Clear	35.9	4.3	42.0	4.2	48.2	4.3	46.8	4.4	37.3	3.8	27.2	3.2

test whether the various audiovisual pairings produce different functions when the percentage of /b/ responses perceived is plotted as a function of different delays.

Neither of these options is entirely satisfactory but we pursued the second option here—assessing differences in the perception of the different audiovisual combinations as a function of delay. The first option has the problem that contradictory information is introduced within the auditory modality. The overall duration of the synthetic or edited stimuli would act as a cue to one speaking rate while the dynamics of the formant change would suggest another speaking rate. The problem with the second option is that it does not offer an unambiguous test of the hypothesis. If functions plotting the percentage of /b/ responses as a function of delay (e.g., Figure 1) for all of the different audiovisual pairings were the same shape, it would suggest that the onset and duration differences were not important. However, if the functions differed in shape or slope, this might be due to either the differences in onset timing and duration or the dynamics themselves interacting with the delays.

In this experiment, we used a subset of the audiovisual speaking style combinations and timing conditions used in the first two experiments. In particular, we used the fast and clear productions and delayed the auditory stimuli only with respect to the video stimuli. We will examine how the slope of the delay function was influenced by the relative speaking rates of the auditory and visual stimuli.

Method

Subjects. Twenty-two undergraduates at Queen's University served as subjects. The subjects were native speakers of Canadian English and reported no speech, language, or hearing problems. All had normal or corrected-to-normal vision. Four subjects responded /b/ for all conditions and were eliminated from the analyses. Thus, the statistical analyses were performed on the data from 18 subjects.

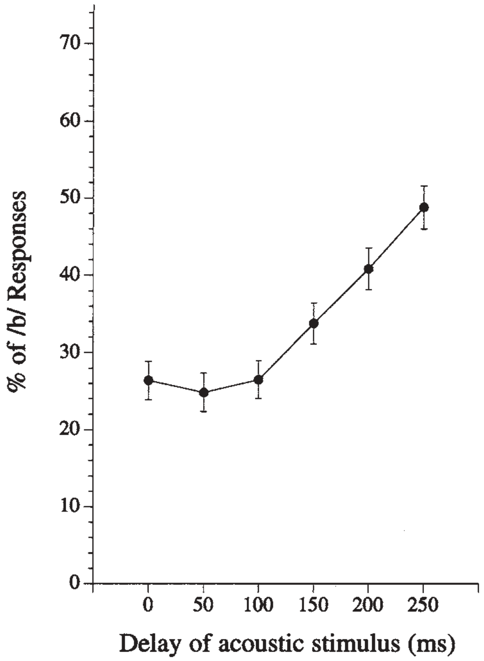

Figure 5. Experiment 3: The percentage of /b/ responses as a function of the delay of the auditory stimuli. The error bars show the standard errors of the means.

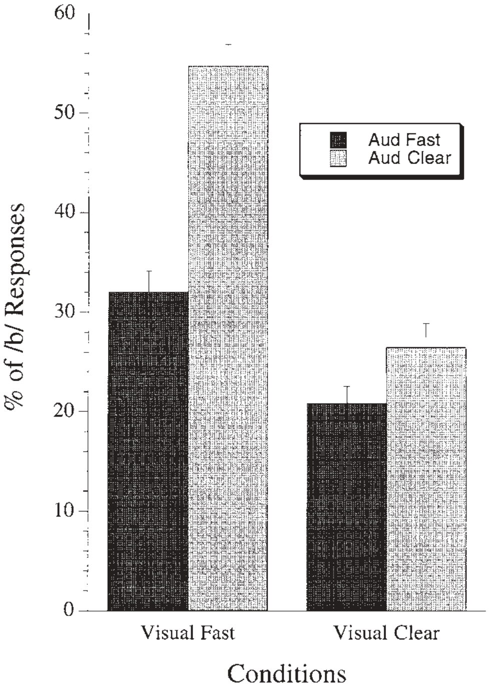

Figure 6. Experiment 3: The percentage of /b/ responses as a function of the different visual and auditory speaking condition combinations. The error bars show the standard errors of the means.

Stimulus materials. Two visual and two auditory stimuli produced by Speakers M.J. and P.B. in Experiment 2 were used in this experiment. Within a speaker, the visual fast and visual clear stimuli and auditory fast and auditory clear stimuli were paired so that each speaker's visual speaking rate utterance was presented with each of the auditory speaking rate utterances and vice versa. This produced 8 pairings (4 per speaker) in which the onsets of the release bursts in the stops were synchronized. In addition, the auditory stimuli were lagged by 50, 100, 150, 200, and 250 msec relative to the timing of the onset of the release burst for the /g/ in the original sound track. This produced 48 audiovisual stimuli in all (eight audiovisual rate pairings × six auditory timing conditions). The blocks of 48 stimuli were shown to the subjects in five different randomized orders.

Results and Discussion

In general, the subjects responded in a fashion similar to that in the first two experiments. As in Experiment 1, there was an overall effect for delay [$F(5,85) = 21.92$, $p < .001$] with the percentage of /b/ responses increasing as the delay increased (Figure 5). As in Experiment 2, there were reliable main effects for visual speaking condition [$F(1,17) = 64.0$, $p < .001$] and auditory speaking condition [$F(1,17) = 32.20$, $p < .001$]. There was also a visual speaking condition × auditory speaking condition interaction [$F(1,17) = 25.40$, $p < .001$]. As can be seen in Figure 6, this interaction is consistent with the concordance effect shown in Experiment 2. A contrast com-

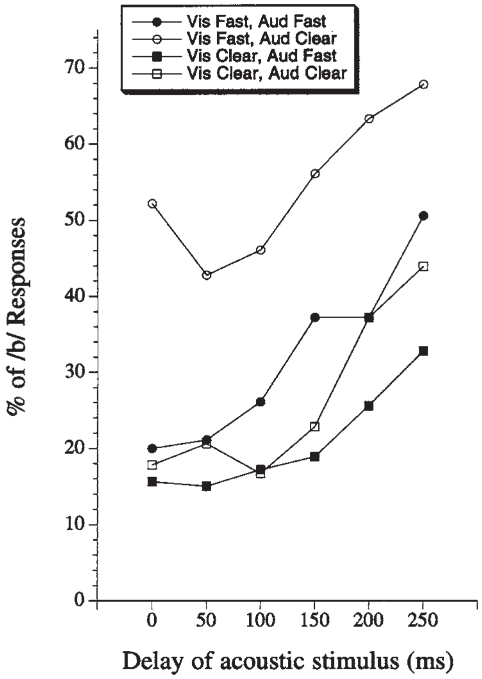

Figure 7. Experiment 3: The percentage of /b/ responses as a function of the delay of the auditory stimuli and the different visual and auditory speaking condition combinations.

paring the matched and unmatched stimuli was reliable [$F(1,17) = 25.40, p < .001$]. The mean percentage of /b/ responses for the matched conditions (visual fast/auditory fast, visual clear/auditory clear) was 29.2%, and for the unmatched conditions (visual fast/auditory clear, visual clear/auditory fast), it was 37.8%. There was no speaker main effect and we will not consider the speaker interactions here.

The main purpose of Experiment 3 was to test whether the desynchrony curves varied as a function of any of the different visual and auditory rate combinations. The delay × visual speaking condition × auditory speaking condition interaction showed no differences in these data [$F(5,85) = 1.97, p > .05$]. The pattern of means for this effect is shown in Figure 7. As can be seen, with the exception of the synchronized mean for the visual fast, auditory clear condition, all of the functions show a similar pattern. Thus, the relative timing of the onsets or offsets of the bisyllables do not seem to have had a large influence on the McGurk effect and thus do not explain the pattern of results observed in the previous experiment.

The mean /d/ and /g/ responses are shown in Table 4. On average, there was a tendency to respond /g/ more than /d/, but the preference was not as extreme as that observed in Experiment 1. The differences between the three experiments in percentages of /d/ and /g/ responses were presumably determined by differences in the individual stimulus characteristics. Analyses of the /d/ and /g/ responses showed reliable delay, visual speaking style, and speaker effects ($p < .01$). In addition, the /d/ responses showed a reliable auditory speaking style effect ($p < .01$).

GENERAL DISCUSSION

The data presented in the three experiments suggest that strict timing of visual and auditory speech information is not the major determinant of audiovisual integration in speech. Subjects' perceptions were influenced by the visual stimuli even when the auditory information lagged the visual information by as much as 180 msec. When the auditory stimuli led the visual stimuli, subjects showed less tolerance for the lack of synchrony. In all three experiments, the data suggested that the dynamic characteristics of articulation affected subjects' perception of the audiovisual stimuli.

The data are consistent with a body of work on the McGurk effect (e.g., Cohen, 1984; Gerdeman, 1994; Massaro & Cohen, 1993; Massaro, Smeele, Cohen, & Sittig, 1995; Tillmann et al., 1984) as well as research on synchrony in normal audiovisual productions (e.g., Dixon & Spitz, 1980; McGrath & Summerfield, 1985; Pandey, Kunov, & Abel, 1986; Smeele, Sittig, & Van Heuven, 1992). This research, similarly, shows that temporal coincidence of information from the auditory and visual channels is not that important. However, in all of this work and in the experiments presented here, the audiovisual stimuli do show some limits on the range over which the signals from the two modalities are treated as

Table 4
Mean Percentage of /d/ and /g/ Responses as a Function of Visual and Auditory Speaking Condition and Delay in Experiment 3

	Visual Fast								Visual Clear							
	Auditory Fast				Auditory Clear				Auditory Fast				Auditory Clear			
	/d/		/g/		/d/		/g/		/d/		/g/		/d/		/g/	
Delay	*M*	*SEM*	*M*	*SEM*	*M*	*SEM*	*M*	*SEM*	*M*	*SEM*	*M*	*SEM*	*M*	*SEM*	*M*	*SEM*
0	40.0	6.0	34.4	6.3	12.2	3.5	32.8	5.1	38.9	6.6	38.9	6.6	27.2	5.4	49.4	6.4
50	36.7	6.0	35.6	6.0	13.4	3.3	38.3	5.4	38.9	6.5	39.4	6.5	26.1	5.6	48.9	6.4
100	38.9	6.3	25.6	4.8	12.8	2.5	34.4	5.3	38.9	6.5	38.3	6.4	32.2	5.7	45.6	6.4
150	27.8	5.2	28.9	5.0	8.9	2.6	30.0	4.9	38.9	6.8	26.1	5.7	25.0	4.0	45.0	5.8
200	25.0	5.2	22.2	4.3	6.7	1.8	35.6	5.8	33.9	6.4	37.8	5.4	17.8	4.0	37.8	5.4
250	17.8	3.8	21.1	4.1	3.3	1.7	22.8	4.6	34.4	5.3	26.1	4.4	20.0	4.4	31.7	5.6

synchronous. What determines the boundaries of this range is unclear. One possibility is that it is not the absolute value of the delay, but rather the timing relative to the duration of the syllable. The information for consonants and vowels is spread across the syllable (see, e.g., Liberman et al., 1967; Öhman, 1967) and syllables can vary in overall duration. This possibility is contradicted by the results of Experiment 3. If syllable duration were a determining factor in the tolerance for audiovisual synchrony, we would have expected that the delay function for the visual fast/auditory fast condition would have been different from the delay function for the visual clear/auditory clear condition. The fast conditions would be expected to show an effect for delay sooner since the delays would exceed the duration of the syllable sooner. There was no evidence for any difference in the functions beyond an overall main effect of response level.[5] Another possibility is that the limitation may be external to the particular stimuli, and the observed pattern may indicate something about general temporal factors in speech information processing. The present experiments do not address this issue; however, the asymmetry in the delay function shown in Figure 1 suggests that general perceptual processing constraints play a role in the observed patterns.

The main effects associated with speaking condition reported in Experiments 2 and 3 suggest that subjects extract information about these speaking conditions from both modalities and that rate or speaking style information from both modalities influences the degree of audiovisual integration in a similar way. Green (1987) reported that the subjects' ratings of speaking rate in auditory, visual, and audiovisual presentations did not differ, and our results are consistent with this finding. Although in the present data there seems to have been a greater range of effects due to visual speaking condition, the pattern of change was the same for both modalities. The observance of greater visual effects does not agree with the findings of Welch, DuttonHurt, and Warren (1986). In their study, subjects were influenced more by the auditory rates of flickering bimodal stimuli than by the visual rates. We cannot determine the source of this difference from the present data. As Green (1987) suggested, however, it may simply be that speaking rate and the type of rate measured by Welch et al. differed in terms of their auditory dominance.

In all three experiments, the subjects showed some sensitivity to the concordance of speaking dynamics between the two modalities. If this finding proves reliable, it would have two implications. First, such a finding would support the view that listeners use the time-varying properties of speech for perceptual grouping and phonetic perception (Remez & Rubin, in press). Remez, Rubin, and colleagues have shown that subjects perceive sinewave speech as speechlike in spite of the loss of all of the short-term spectra of natural speech (see, e.g., Remez, Rubin, Berns, Pardo, & Lang, 1994). In sinewave speech, time-varying sinusoids track the formant center frequencies of natural utterances. These stimuli, thus, do not have harmonic structure, fundamental frequency, or normal formant bandwidth. What the sinewave stimuli do provide for listeners is information about the rate of change of the vocal tract shapes. According to Remez and Rubin, this information is sufficient to specify that the sinewave stimuli are speechlike, and it usually permits the identification of the speech stimuli. In the McGurk effect, it may be that the information about the rate of change of the vocal tract is extracted from the stimuli in both modalities. Summerfield (1987), in fact, has suggested that one possible metric for audiovisual integration is the pattern of changes over time in articulation. In his view, a promising possibility is that listeners are sensitive to the dynamics of vocal tract change. Similar proposals have been used by Fowler and Dekle (1991) and Bernstein, Coulter, O'Connell, Eberhardt, and Demorest (1992) to account for subjects' ability to perceive haptic/auditory and haptic/visual speech stimuli. This is not to say that the movement dynamics provide the only information that can aid audiovisual integration in speech. In our experiments, the stimuli in each modality were rich in information. As a result, there were numerous clues to the identity of the tokens. However, it appears that one significant influence might be how well the stimuli match in the information that they provide about the rate of change of the vocal tract.

Other evidence that the dynamics of speech are important in audiovisual communication comes from studies of the minimum video frame rates that allow visual speech perception. Vocal tract movements are relatively slow (<20 Hz), and thus the dynamic facial information has a limited frequency bandwidth. Recent evidence suggests that subjects need video frame rates just fast enough to capture the motions of the face during speech. Vitkovich and Barber (1994) suggested that a frame rate of about 17 Hz may be sufficient for the transmission of facial information. Below this frame rate, intelligibility will suffer for some subjects.

The second implication of the concordance finding is that it would add to our understanding of how the information from the two modalities is combined. Recently, Green et al. (1991) have argued that information about the speaker's voice characteristics is used to normalize speech stimuli *before* the information from the auditory and visual systems is combined. Massaro (1987) has also proposed that auditory and visual information is processed to some degree before integration takes place (cf. Braida, 1991). Our results suggest that dynamic information from the two modalities is available until the point of audiovisual integration. As Miller (1986) stated, the analysis of rate-dependent information seems to be an obligatory part of speech processing, even in audiovisual perception.

The concordance analyses in this paper have limitations, however. It is clear that the results in Experiments 2 and 3 were influenced to a great extent by differences in one or two cells, though it is not clear what the source of the differences was. In Experiment 2, for example, the visual fast/auditory clear cell showed considerably more /b/ responses than average, whereas the visual clear/auditory fast condition produced fewer /b/ responses than average. When averaged together, these "most discrepant" stimuli still yield the highest average rate of /b/ responses, but the effect derives from the visual fast/auditory clear cell. One possibility is that the auditory vowel onset in this combination occurs too early because of the longer duration of the auditory clear stimulus. Even though the burst onset is synchronized, the early vowel onset could disrupt audiovisual integration in a manner similar to the effects shown in Experiment 1 for the conditions in which the auditory stimulus preceded the normal audiovisual timing. In support of this hypothesis is the fact that the visual fast/auditory clear condition in Experiment 3 was most discrepant at synchrony (Figure 7). When the acoustics are delayed and thus the auditory vowel onset is later, the desynchrony function becomes similar in shape to that produced by the other audiovisual pairings. A second possibility is that information from a single modality is subject to threshold effects in audiovisual communication. If the information provided by a given modality is below a certain threshold and is presented along with strong information in another modality, the strong modality may dominate. This would mean that the visual fast stimuli were simply too weak when paired with the auditory clear stimuli. This issue cannot be resolved with the current data and will require additional research.

Finally, we note that we still know relatively little about the time course of audiovisual information processing in speech perception. Smeele, Sittig, and Van Heuven (1994; see also Green & Gerdeman, 1995) have shown that subjects have access to visual information about place of articulation earlier than the auditory information. In part this is because there is significant visual motion prior to the auditory onset of the syllable. Information such as this on the time course of the processing of audiovisual perception will be important for understanding the nature of the integration of information from different modalities.

In summary, the present experiments indicate that the relative timing of visual and auditory information is not critical in speech perception. On the other hand, the dynamics of articulation within each of the modalities may be influencing audiovisual perception. The results support the view that rate-dependent information is fundamental to phonetic perception.

REFERENCES

Abry, C., & Boë, L. J. (1986). Laws for lips. *Speech Communication*, **5**, 97-193.

Bernstein, L. E., Coulter, D., O'Connell, M. P., Eberhardt, S., & Demorest, M. (1992, June). *Vibrotactile and haptic speech codes*. Paper presented at the Second International Conference on Tactile Aids, Hearing Aids, & Cochlear Implants, Stockholm.

Bernstein, L. E., & Eberhardt, S. (1986). *Audio–visual stimuli*. Johns Hopkins University, Department of Electrical and Computer Engineering.

Braida, D. L. (1991). Crossmodal integration in the identification of consonant segments. *Quarterly Journal of Experimental Psychology*, **43**, 647-677.

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

Dixon, N., & Spitz, L. (1980). The detection of audiovisual desynchrony. *Perception*, **9**, 719-721.

Dunnett, C. W. (1955). A multiple comparison procedure for comparing several treatment means with a control. *Journal of the American Statistical Association*, **50**, 1096-1121.

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.

Gagne, J.-P., Masterson, V., Munhall, K. G., Bilida, N., & Querengesser, C. (1994). Across talker variability in speech intelligibility for conversational and clear speech: A crossmodal investigation. *Journal of the Academy of Rehabilitative Audiology*, **27**, 133-158.

Gay, T. (1978). Effect of speaking rate on vowel formant transitions. *Journal of the Acoustical Society of America*, **63**, 223-230.

Gay, T. (1981). Mechanisms of the control of speech rate. *Phonetica*, **38**, 148-158.

Gerdeman, A. (1994). *Temporal incongruity and the McGurk effect*. Unpublished master's thesis, University of Arizona, Tucson.

Gracco, V., & Abbs, J. (1986). Variant and invariant characteristics of speech movements. *Experimental Brain Research*, **65**, 156-166.

Green, K. P. (1987). The perception of speaking rate using visual information from a talker's face. *Perception & Psychophysics*, **42**, 587-593.

Green, K. P., & Gerdeman, A. (1995). Cross-modal discrepancies in coarticulation and the integration of speech information: The McGurk effect with mismatched vowels. *Journal of Experimental Psychology: Human Perception & Performance*, **21**, 1409-1426.

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, N. A. (1988, November). *Factors affecting the integration of auditory and visual information in speech: The effect of vowel environment*. Paper presented at the annual meeting of the Acoustical Society of America, Honolulu.

Green, K. P., Kuhl, P. K., Meltzoff, A. N., & Stevens, E. B. (1991). Integrating speech information across talkers, gender, and sensory modality: Female faces and male voices in the McGurk effect. *Perception & Psychophysics*, **50**, 524-536.

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

Green, K. P., Stevens, E. B., & Kuhl, P. K. (1994). Talker continuity and the use of rate information during phonetic perception. *Perception & Psychophysics*, **55**, 249-260.

Hirsh, I. J., & Sherrick, C. E. (1961). Perceived order in different sense modalities. *Journal of Experimental Psychology*, **62**, 423-432.

Kashino, M., & Craig, C. (1994). The influence of knowledge and experience during the processing of spoken words: Non-native listeners. In *Proceedings of the International Conference on Spoken Language Processing* (pp. 2047-2050).

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

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H and H theory. In W. Hardcastle & A. Marchal (Eds.), *Speech production and speech modeling* (pp. 403-439). Dordrecht: Kluwer.

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

Manuel, S. Y., Repp, B., Studdert-Kennedy, M., & Liberman, A. (1983). Exploring the "McGurk effect." *Journal of the Acoustical Society of America*, **74**, S66.

Massaro, D. W. (1987). *Speech perception by ear and eye*. Hillsdale, NJ: Erlbaum.

Massaro, D. W., & Cohen, M. M. (1993). Perceiving asynchronous bimodal speech in consonant-vowel and vowel syllables. *Speech Communication*, **13**, 127-134.

Massaro, D. W., Smeele, P. M. T., Cohen, M. M., & Sittig, A. C.

(1995). *Perception of asynchronous and conflicting visual and auditory speech*. Manuscript submitted for publication.

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 speech. *Nature*, **264**, 746-748.

Miller, J. (1986). Rate-dependent processing in speech perception. In A. Ellis (Ed.), *Progress in the psychology of language* (Vol. 3, pp. 119-157). Hillsdale, NJ: Erlbaum.

Miller, J., & Baer, T. (1983). Some effects of speaking rate on the production of /b/ and /w/. *Journal of the Acoustical Society of America*, **73**, 1751-1755.

Öhman, S. (1967). Numerical model of coarticulation. *Journal of the Acoustical Society of America*, **41**, 310-320.

Pandey, C. P., Kunov, H., & Abel, M. S. (1986). Disruptive effects of auditory signal delay on speech perception with lip-reading. *Journal of Auditory Research*, **26**, 27-41.

Picheny, M. A., Durlach, N., & Braida, L. (1985). Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech. *Journal of Speech & Hearing Research*, **28**, 96-103.

Pick, H., Warren, D., & Hay, J. (1969). Sensory conflict in judgments of spatial direction. *Perception & Psychophysics*, **6**, 203-205.

Remez, R. E., & Rubin, P. E. (in press). Acoustic shards, perceptual glue. In J. Charles-Luce, P. A. Luce, & J. R. Sawusch (Eds.), *Theories in spoken language: Perception, production, and development*. Norwood, NJ: Ablex.

Remez, R. E., Rubin, P. E., Berns, S. M., Pardo, J. S., & Lang, J. M. (1994). On the perceptual organization of speech. *Psychological Review*, **101**, 129-156.

Scheirman, G. L., & Cheetham, P. J. (1990). Motion measurement using the Peak Performance Technologies system. *Society of Photo-optical Instrumentation Engineers Proceedings*, **1356**, 67-70.

Sekiyama, K., & Tohkura, Y. (1991). McGurk effect in non-English listeners: Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility. *Journal of the Acoustical Society of America*, **90**, 1797-1805.

Smeele, P. M. T., Sittig, A. C., & Van Heuven, V. J. (1992). Intelligibility of audio-visually desynchronized speech: Asymmetrical effect of phoneme position. In *Proceedings of the International Conference on Spoken Language Processing* (pp. 65-68).

Smeele, P. M. T., Sittig, A. C., & Van Heuven, V. J. (1994). Temporal organization of bimodal speech information. In *Proceedings of the International Conference on Spoken Language Processing* (pp. 1431-1434).

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. (1992). Lipreading and audio-visual speech perception. *Philosophical Transactions of the Royal Society of London: Series B*, **335**, 71-78.

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

Tillmann, H. G., Pompino-Marschall, B., & Porzig, H. (1984). Zum Einfluß visuell dargeborener Sprachbewegungen auf die Wahrnehmung der akustisch kodierten Artikulation. *Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation der Universität München*, **19**, 318-338.

Vatikiotis-Bateson, E., Eigsti, I., & Yano, S. (1994). Listener eye movement behavior during audiovisual speech perception. In *Proceedings of the International Conference on Spoken Language Processing* (pp. 527-530).

Vitkovich, M., & Barber, P. (1994). Effects of video frame rate on subjects' ability to shadow one of two competing verbal passages. *Journal of Speech & Hearing Research*, **37**, 1204-1210.

Ward, M. (1992). *The effect of auditory–visual dysynchrony on the integration of auditory and visual information in speech perception*. Unpublished bachelor's thesis, Queen's University, Kingston, Ontario.

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

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

NOTES

1. It was reasoned that a response of /b/ indicated that the visual stimuli had no influence on the subject's judgment. Any non-/b/ response could be interpreted as being caused by the different visual conditions. Although in general a non-/b/ response could be caused by an error in auditory perception, this cannot account for any systematic differences between conditions since the acoustic stimuli were held constant across conditions. Thus the relative number of /b/ responses between conditions rather than the absolute number indicates the visual influence.

2. Some caution should be exercised in interpreting these trajectories. The measures are only gross estimates of oral aperture since the measures contain some amount of head motion and only the aperture height is being measured. In addition, the amount of movement is presumably only one of the determinants of visual intelligibility.

3. We have not interpreted the /d/ response as a fusion response and the /g/ as a visual dominance response in this experiment or elsewhere in the paper because we believe that this interpretation is not always well founded. For many speakers, /d/ and /g/ are visually indistinguishable, and for some stimuli, visual /g/s are consistently labeled /d/. The interpretation of /d/ as a fusion response and /g/ as a visual dominance response is not warranted without independent psychophysics on the visual and auditory stimuli. A predominance of /d/ could reflect the predominance of /d/s in the English lexicon, the most frequent response to the visual stimuli, a fusion response, and so on.

4. The clear condition produced longer durations but also larger movements. Thus the velocity of the facial movement was higher than for the other two conditions for the clear condition. The patterns shown by the other 2 speakers were less clear. L.J. showed the smallest movements and little difference across conditions. P.B. showed more movement in the fast condition than in the other two conditions.

5. It may be that the three speaking conditions that we used here did not differ greatly in the duration of the critical information about the moving vocal tract. Changes in rate were not uniformly distributed across syllables.

(Manuscript received December 15, 1994;
revision accepted for publication July 26, 1995.)