TR－H－112

Temporal constraints on the perception of the McGurk effect

K.G. Munhall
P. Gribble (McGill University)
L. Sacco (Queen's University)
M. Ward (Queen's University)

1994. 12. 12

ATR人間情報通信研究所

〒619-02 京都府相楽郡精華町光台2-2　☎07749-5-1011

ATR Human Information Processing Research Laboratories

2-2, Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-02 Japan

Telephone: +81-7749-5-1011
Facsimile: +81-7749-5-1008

Temporal constraints on the perception of the McGurk effect.

K.G. Munhall (Queen's University & ATR Human Information Processing Research Laboratories)

P.Gribble (McGill University)

L.Sacco and M. Ward (Queen's University)

Submitted to: Perception and Psychophysics

Abstract

Three experiments are reported on the influence of different timing relations on the perception of 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 acoustic stimuli lagged the visual stimuli by as much as 180 ms. In addition, a stronger McGurk effect was found when the visual and acoustic 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 both the visual and auditory speaking-condition manipulations independently influenced perception. In addition, the conditions in which the auditory and visual stimuli were spoken at the same rate showed a stronger McGurk effect. In the third experiment we combined auditory and visual stimuli produced at different speaking rates 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 rate combinations. The results indicate that perceivers are sensitive to concordance of the time-varying aspects of speech but they do not require temporal coincidence of that information.

Temporal constraints on the perception of the McGurk effect.

K.G. Munhall (Queen's University & ATR Laboratories)

P.Gribble (McGill University)

L.Sacco and M. Ward (Queen's University)

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 lip-reading as well as studies of speech in noise (e.g., Sumby and 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 acoustic 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 acoustic stimulus is /bi/, subjects report hearing /di/. This phenomenon has been called the McGurk effect (McGurk and MacDonald, 1976) and in this paper we use this effect to study audiovisual speech perception.

Since the original report on the McGurk effect (McGurk and 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 the McGurk effect 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 acoustic perception of voicing (Green & Kuhl, 1989), etc. However, as Green et al., (1991) point out, these papers have not described the basic 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 badly dubbed foreign-language movies to quickly notice the disparity between 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 acoustic speech signals. It is clear that for brief, nonverbal stimuli, people are very sensitive to intermodal timing (e.g. Hirsh & Sherrick, 1961) with timing differences of less than 20 ms 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 ms before subjects noticed the discrepancy. Similar time values were found by Koenig (1965, cited in McGrath and 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 ms. However, there was a reliable decrease in transcription performance when the audio was delayed 160 ms.

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 will always vary considerably. However, what is clear from the existing data is that the delays required to disrupt speech perception are surprisingly large. While a great deal of evidence from the study of the acoustic 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 demi-syllable range. This fact has important practical and theoretical implications (McGrath and 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 occurred 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 that 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

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 Eberhard (1986) database. In Experiments 2 and 3 the images were from a videodisc recorded at Queen's University. The acoustic 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 inch 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 a MG Electronics Cabaret speaker that was placed directly below the monitor. Custom software was used to control the videodisc trials, play the acoustics synchronously with the video, and record subjects'

responses from the keyboard. The software allowed the acoustics to be flexibly timed with approximately 1 ms accuracy across trials.

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 acoustic stimulus, /b/. This timing relation was considered synchronous and the experimental software allowed this timing relationship to be reliably reproduced.

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 speakers and to listen to the acoustic output from the speaker and report what the 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 3 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 two seconds following the subject's response. Between trials the screen was blackened.

EXPERIMENT I

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 & Cohen (1993) manipulated temporal asynchrony to study how visual /bɑ/ and acoustic /dɑ/ are combined to be perceived as /bdɑ/. Subjects perceived /bdɑ/ even when the acoustic /dɑ/ preceded the visual /bɑ/ by as much as 200 ms. 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. Tillman, Pompino-Marschall and Porzig (1984) have shown that for German subjects the combination of a visual "gier" and acoustic "bier" produces the McGurk percept "dier" across a wide range of temporal alignments. Tillman et al. (1984) varied the temporal alignment of the acoustics by

±500 ms. They reported that /d/ responses exceeded /b/ responses over a wide range of values (± 250 ms). More recently, Ward (1992) reported that acoustic delays of up to 300 ms still produced a significant number of McGurk responses.

The present study aims to replicate the study of Tillman 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 2 different patterns of visual motion.

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 fourth 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 Eberhard (1986) videodiscs. The acoustic stimulus was a digitized version of the same speaker's productions of /ɑbɑ/. The timing of the acoustics varied in 60 ms steps from 360 ms prior to synchrony to 360 ms after synchrony. Thus, there were 13 audiovisual pairings for each vowel context.

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 is.[1] A 2-way repeated measures ANOVA (audiovisual synchrony X vowel) was used to analyze the data. The overall results are plotted in Figure 1. As can be seen the vowel and synchrony conditions influence 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/ response 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/ response.

The vowel effect is opposite to results reported by Green, Kuhl, & 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 is paired with an acoustic /ɑ/ stimulus. Thus, we cannot determine if the source of the difference is 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 speakers differ greatly in the pattern of this intelligibility across different words. In part, these differences are casued by the amount of movement for a given 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 the intervocalic consonant , /g/, and opens for the second vowel. Finally, the mouth closes following

[1]It was reasoned that a response of /b/ indicated that the visual stimuli had no influence on the subject's judgement. Any non-/b/ response could be interpretted as being caused by the different visual conditons. While 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 rather than the absolute number indicates the visual influence.

the end of the bisyllable. As can be seen, the relative amount of visual motion is much less for the intervocalic /g/ in the /i/ context in the stimuli used in this experiment.[2]

The second 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 acoustic changes to get strong audiovisual fusions. This will be explored directly in Experiment II.

The synchrony manipulation produced a V-shaped function for the rate of /b/ response. 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 X 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 is a tendency to respond /b/ more frequently when the audio signal leads the video than vice versa ($F(1,168)=60.54$, $p<.001$). In fact the lowest rate of /b/ response occurs when the audio lags the video by 60 ms rather than when the audio signal is 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 was 30 meters away from the person they were speaking to, the acoustics would reach the listener about 88 ms after sight of the corresponding facial movements. Smeele, Sittig, & van Heuven (1992) and Dixon & Spitz (1980) have reported similar asynchronies to the one we have observed.

---

[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.

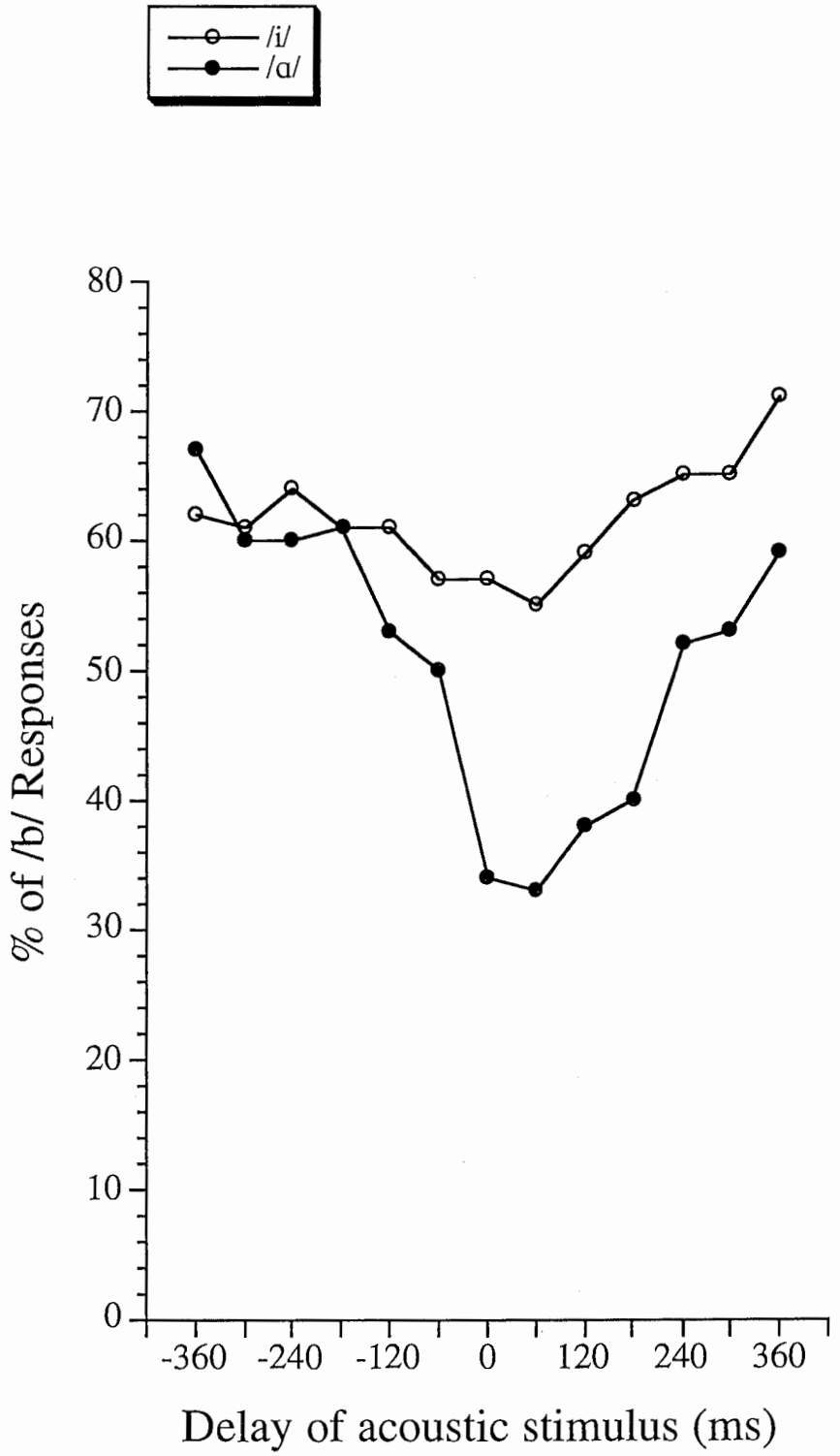

Figure 1 The percentage of /b/ responses as a function of the delay of the acoustic stimuli. Negative numbers on the abscissa indicate that the acoustic stimulus preceded the visual stimulus. Data for the two vowel contexts are plotted separately.

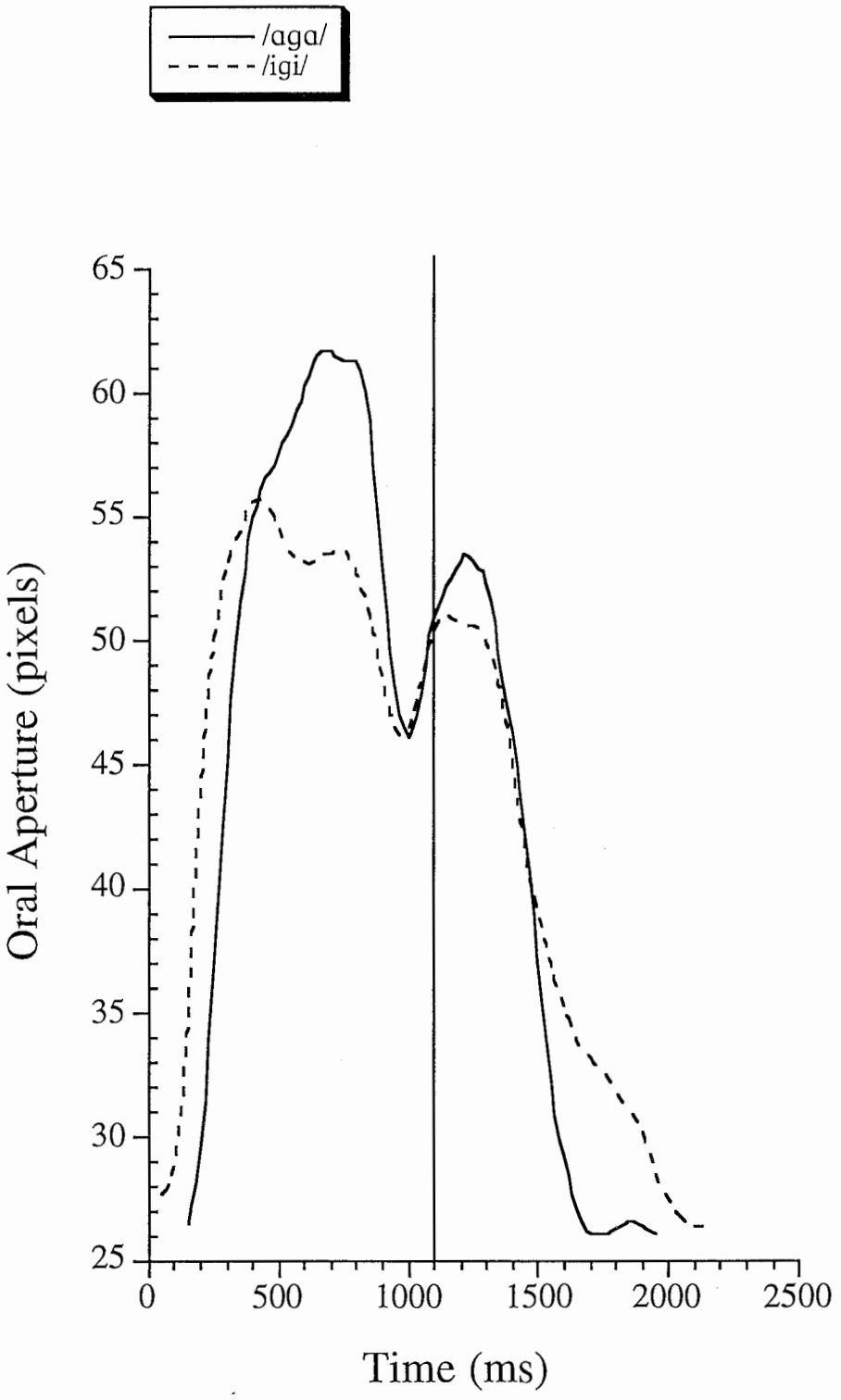

Figure 2. Kinematics of the oral aperture in the visual stimuli used in Experiment I. Data for the two vowel contexts are plotted separately. The traces are lined up at the onset of the acoustic burst.

The overall pattern of results are consistent with the data of Cohen (1984), Tillman et al. (1984), Ward (1992), Massaro & Cohen (1993) and Green (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 percent of /b/ responses was reliably higher ($p<.05$) than the zero lag condition when the acoustics preceded the visual timing by 60 ms and when the acoustics followed the visual timing by 240 ms for the vowel /ɑ/. The data for the vowel /i/ were not examined because the function was so flat.

Two final aspects of the data warrant comment. As can be seen in Figure 1 the subjects never show 100% /b/ response even when the audio signals are 360 ms 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 because auditory-only tests under these conditions in the same laboratory have shown very high rates of /b/ response. Further, Tillman 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 are 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 & 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 while others showed little effect.

EXPERIMENT II

In Experiment I, synchrony was defined with respect to a particular moment, the onset of the release burst. While this is an important point in time for stop consonant production and perception, the information for the stop is not localized at any single point in time (e.g., Kashino & Craig, 1994). 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. The possibility arose in Experiment I that audiovisual integration was greater when information in the two modalities was consistent. In Experiment II we investigate the use of this dynamic information. To this end, we manipulated speaking rate acoustically 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 and Baer, 1980). Others have shown that the rate of visual speech information can influence the perception of acoustic speech categories. Green and Miller (1985), for example, showed that the perceived boundary along a continuum of acoustic 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).

SUBJECTS

30 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 produce in 3 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 influenced the strength of the McGurk effect. The visual stimuli were /ɑgɑ/ 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 other effects on the speech than simply a change in rate (e.g. Lindblom, 1990; Picheny, Durlach, & Braida, 1985), however it allowed us to naturally produce an utterance with a longer duration without resorting to the highly unnatural instruction "speak slowly". The average acoustic durations across speakers of the /ɑgɑ/ utterances used for visual stimuli were 304.85, 399.12, and 491.44 ms for fast, normal and clear respectively.

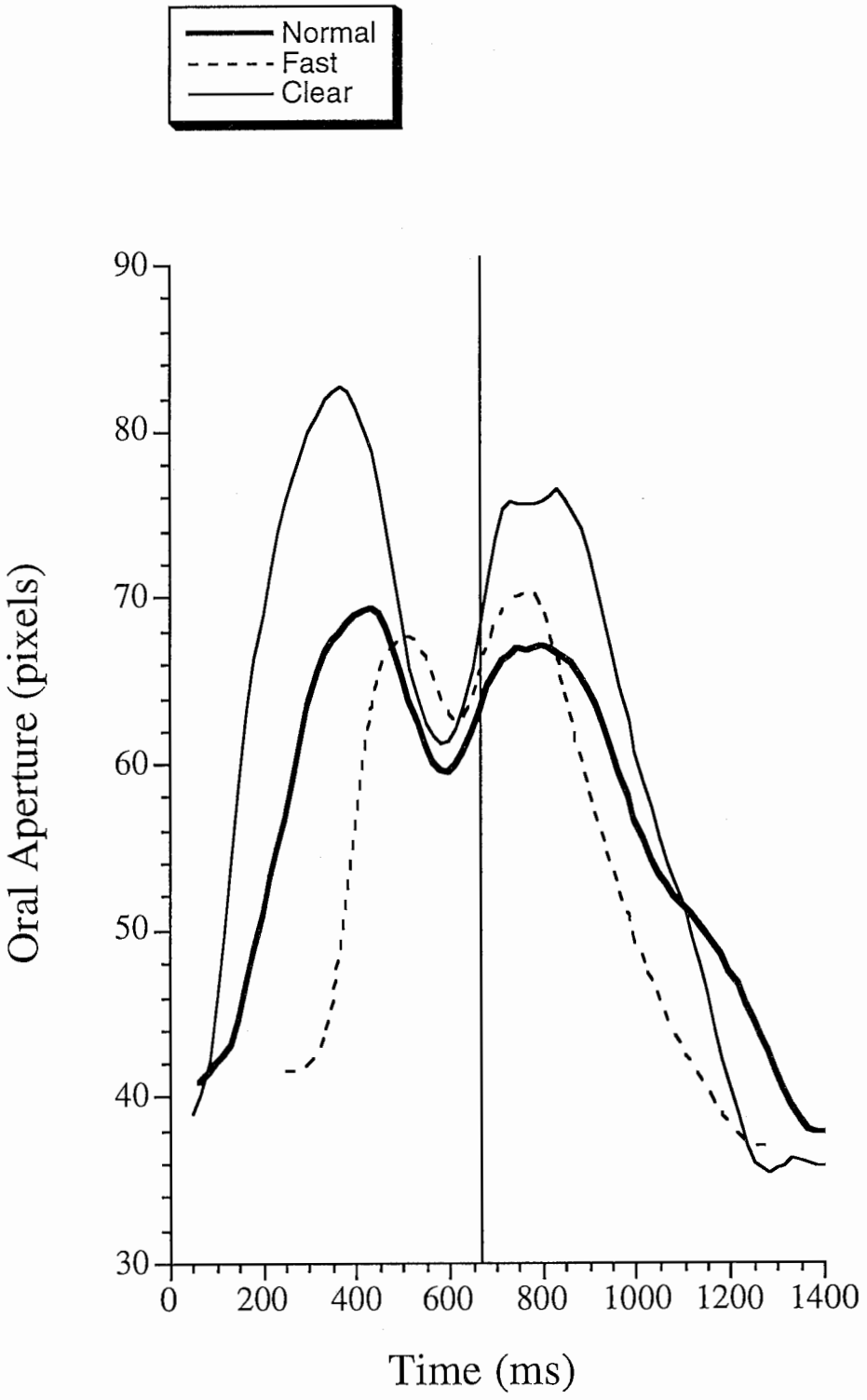

Figure 3. Kinematics of the oral aperture for the visual stimuli for speaker MJ used in Experiment II. Data for the three speaking rates are plotted separately. The traces are lined up at the onset of the acoustic burst.

The acoustic stimuli were productions of /ɑbɑ/ produced by the three speakers in the same three speaking conditions. The average durations across speakers of the acoustic /ɑbɑ/ stimuli were 302.89, 378.18, and 488.15 ms for fast, normal, and clear respectively. Figure 3 shows the kinematic patterns for the motion of the oral aperture for speaker MJ. The traces plot the motion of the mouth from a closed position throught 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[3]. 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 such that the utterance for each visual speaking condition was presented with the utterance for each acoustic speaking condition. This produced nine pairings for each of the three speakers and thus 27 audiovisual stimuli in all. All of the acoustic 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 dependent variable was the percentage of /b/ responses. The data were analyzed in a 3-way, repeated measures ANOVA (speaker X visual speaking condition X audio speaking condition). The three speakers differed in the percentage of /b/ responses that they elicited ($F(2, 58) = 10.27$, $p<.01$). As can be seen in Table 1, speaker MJ's stimuli yielded the least /b/ responses, while speaker LJ produced the

---

[3] The clear condition produced longer durations but also larger movements. Thus the velocity of the facial movement is higher for this condition. The patterns shown by the other two speakers were less clear. LJ showed the smallest movements and little difference across conditions. PB showed more movement in the fast condition than the other two conditions.

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 moves from fast to normal to clear, the information within a modality increases in influence. In the auditory channel, this manifests itself as an increased number of /b/ responses. The auditory fast rate produced an overall average of 17.22 % /b/ responses while the clear condition produced 27.96%. In the visual channel, this manifests itself as a decreased number of /b/ response. The visual fast condition produced an overall average of 30.52% /b/ responses while the clear condition produced 16.26% /b/ responses. (See Figure 4.)

If the concordance between the information from the two modalities is important for producing audiovisual integration, then we would expect that there should be a visual speaking condition X 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 X 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 averages of the matched and unmatched audiovisual conditions differed ($F(1,116) = 15.62$, $p<.01$) with the average of the three matched conditions producing, on average, less /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 percent of /b/ responses for the matched (vis. fast/aud. fast, vis. normal/aud. normal, vis. clear/aud. clear), discordant (vis. fast/aud. normal, vis. normal/aud. fast, vis. normal/aud. clear, vis. clear/aud. normal) and most discordant conditions (vis.

fast/aud. clear, vis. clear/aud. fast) were 19.41, 20.42, and 27.28 respectively. Thus,

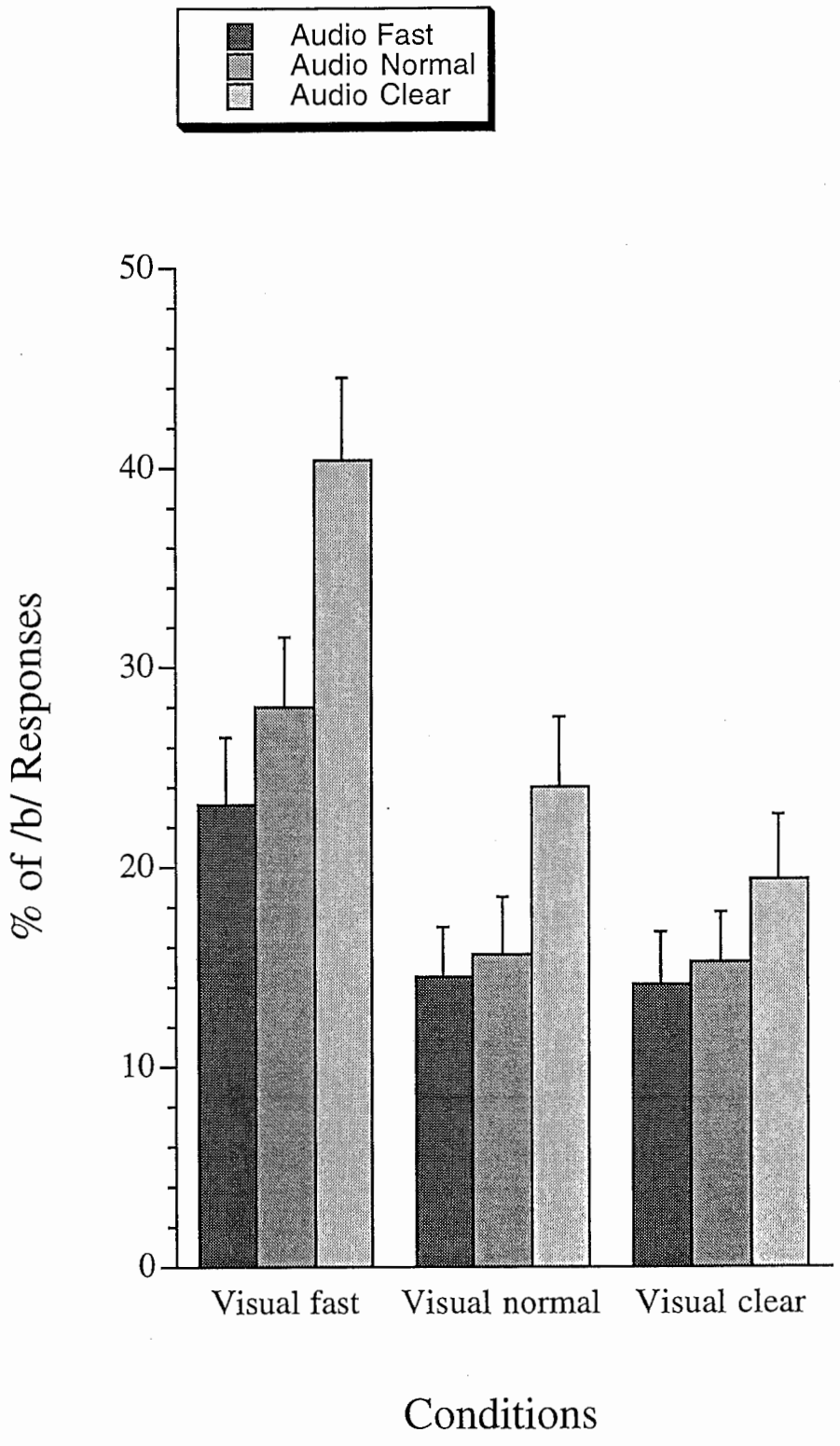

Figure 4. The percentage of /b/ responses as a function of the different visual and auditory speaking condition combinations. The error bars show the standard errors of the means.

Table 1

	Speaker								
	MJ			PB			LJ		
Visual Rate	Fast	Normal	Clear	Fast	Normal	Clear	Fast	Normal	Clear
Auditory Rate									
Fast	16.0	11.3	10.0	16.7	10.0	6.7	36.7	22.0	25.7
	(4.8)	(3.8)	(3.8)	(5.0)	(3.1)	(2.4)	(6.9)	(6.0)	(6.2)
Normal	23.3	11.0	9.0	30.7	18.0	17.3	30.0	18.0	19.3
	(5.2)	(4.0)	(3.2)	(6.5)	(5.8)	(4.7)	(6.7)	(5.0)	(5.1)
Clear	22.0	9.0	5.0	37.0	24.0	12.3	62.3	39.0	41.0
	(5.3)	(4.4)	(3.0)	(6.2)	(5.3)	(3.3)	(7.8)	(7.2)	(7.3)

Mean percentage of /b/ responses as a function of speaker, visual speaking condition and auditory speaking condition. The value in parentheses below each mean is the standard error of the mean.

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 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. While this effect of visual concordance is reliable, we note that it is not large. The two contrasts testing the concordance effect account 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 since we know little about the characteristics of individual speakers that make them more or less intelligible (cf. Gagne et al., 1994) and because they did not seem to change any of the major patterns. The speaker X visual speaking condition X auditory speaking condition interaction was not reliable ($F(8,232) = 1.44$, $p>.1$). The speaker X speaking condition interaction was significant ($F(4,116) = 10.85$, $p<.01$). The speaker X visual speaking condition interaction was also reliable ($F(4,116) = 2.72$, $p<.05$).

As a second test of the concordance hypothesis, we examined the percent of /d/ responses. The rationale was that /d/ responses indicated a more obvious integration of the auditory and visual information. A /g/ responses could be interpreted as visual dominance rather than a response determined by both modalities. The analysis showed reliable speaker, visual speaking condition, and auditory speaking condition effects but the visual speaking condition X auditory speaking condition interaction is what concerns us. There was a reliable interaction ($F(4,116)=3.07$, $p<.01$) and the orthogonal contrasts showed patterns similar to the /b/ results. The averages of the

matched and unmatched audiovisual conditions differed ($F(1,116) = 4.77$, $p<.05$) with the average of the three matched conditions producing, on average, more /d/ responses. The average of the two conditions that were most discordant (visual fast/auditory clear; visual clear/auditory fast) produced a lower rate of /d/ responses than the average of the unmatched conditions that were closer to each other in rate ($F(1, 116) = 19.86$, $p<.001$). The mean percent of /d/ responses for the matched, discordant and most discordant conditions were 42.7, 42.3, and 37.4 respectively.

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 III

In Experiment II we maintained synchrony at the point of acoustic release while we have 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 also were the most discordant in terms of overall duration and timing of the onsets and offsets. There are two ways that this 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 I for different audiovisual rates. By doing this we could test whether the various audiovisual pairings produce different functions for the perception of the McGurk effect at different delays. Neither of these options are entirely satisfactory but we will pursue the second option here. The first option has the problem that contradictory information is introduced within the acoustic 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 percent of /b/ responses as a function of delay (e.g., Fig. 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 may be due to either the differences in onset timing and duration or the dynamics themselves interacting with the delays.

In this experiment we use a subset of the audiovisual rate combinations and timing conditions only where the acoustics are delayed with respect to the video stimuli. We will examine how slope of the delay function is influenced by the relative speaking rates of the auditory and visual stimuli.

SUBJECTS

22 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.

STIMULUS MATERIALS

Two visual and two acoustic stimuli produced by Speakers MJ and PB in Experiment II were used in this experiment. Within a speaker, the visual fast and visual clear stimuli and acoustic fast and acoustic clear stimuli were paired such that each speaker's visual-speaking rate utterance was presented with each of their acoustic-speaking rate utterances and vise versa. This produced 8 pairings (4/speaker) in which the onsets of the release bursts in the stops were synchronized. In addition, the acoustics were lagged by 50, 100, 150, 200 and 250 ms 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 (8 audiovisual rate pairings X 6 acoustic timing conditions). The blocks of 48 stimuli were shown to the subjects in 5 different randomized orders.

RESULTS and DISCUSSION

In general, the subjects responded in a fashion similar to the first 2 experiments. As in Experiment I, 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. (See Figure 5.) As in Experiment II, there were reliable visual rate ($F(1,17)=64.0$, $p<.001$) and auditory rate ($(1,17)=32.20$, $p<.001$) main effects. There was also a visual rate X auditory rate 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 II. A contrast comparing the matched and unmatched stimuli was reliable ($F(1,17)= 25.40$, $p<.001$). The mean percent of /b/ responses for the matched conditions (vis. fast/aud. fast, vis. clear/aud. clear) was 29.2 and for the unmatched conditions (vis. fast/aud. clear, vis. clear/aud.fast) was 37.8. There was no speaker main effect and we will not consider the speaker interactions here.

The main purpose of Experiment III was to test whether the desynchrony curves varied as a function of any of the different visual and auditory rate combinations. The delay X visual rate X auditory rate interaction showed no differences in these data ($F(5,85)=1.97$, $p>.05$). The pattern of means for this effect are 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 a large influence on the McGurk effect and thus are not the explanation for the pattern of results observed in the previous experiment. As in Experiment II there is a tendency for the congruent auditory and visual stimuli to exhibit more McGurk effects.

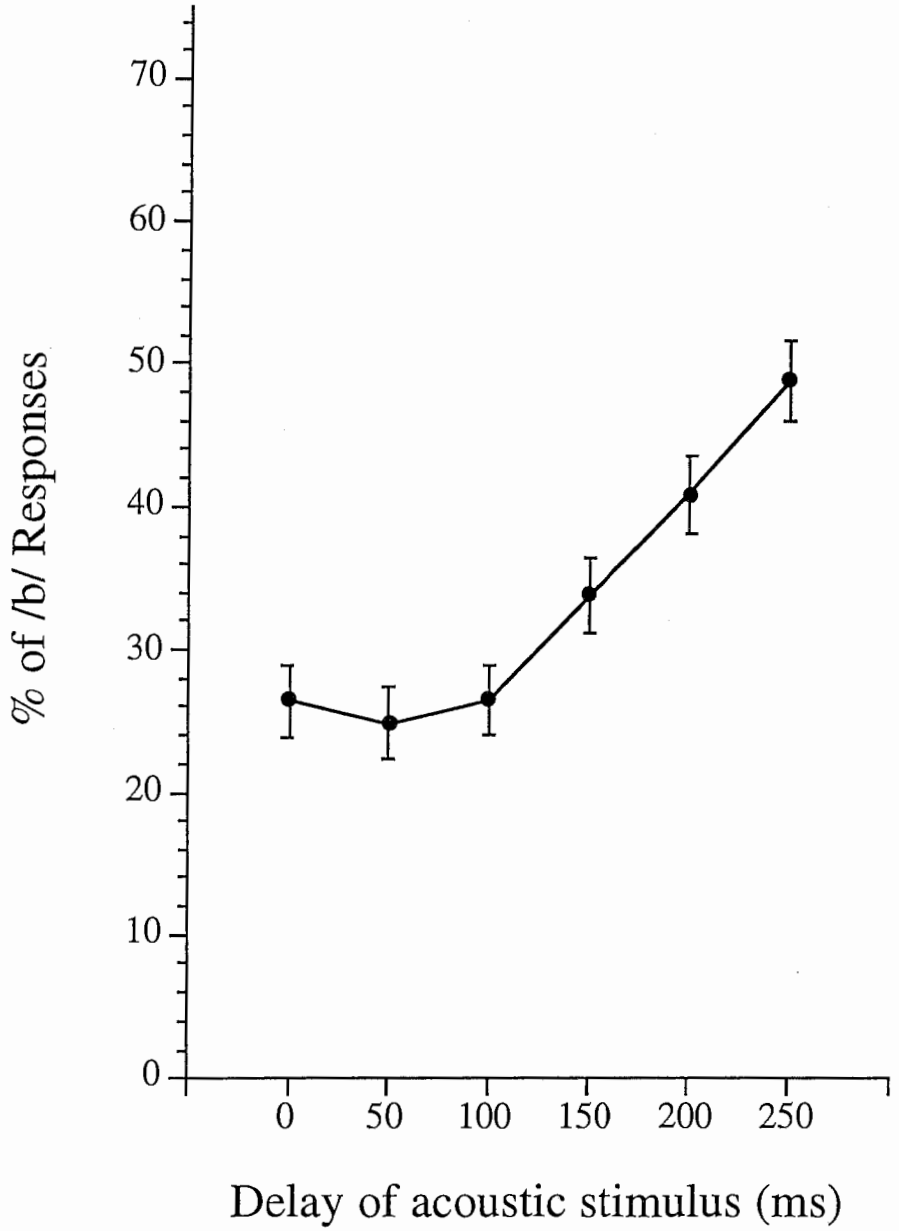

Figure 5. The percentage of /b/ responses as a function of the delay of the acoustic stimuli. The error bars show the standard errors of the means.

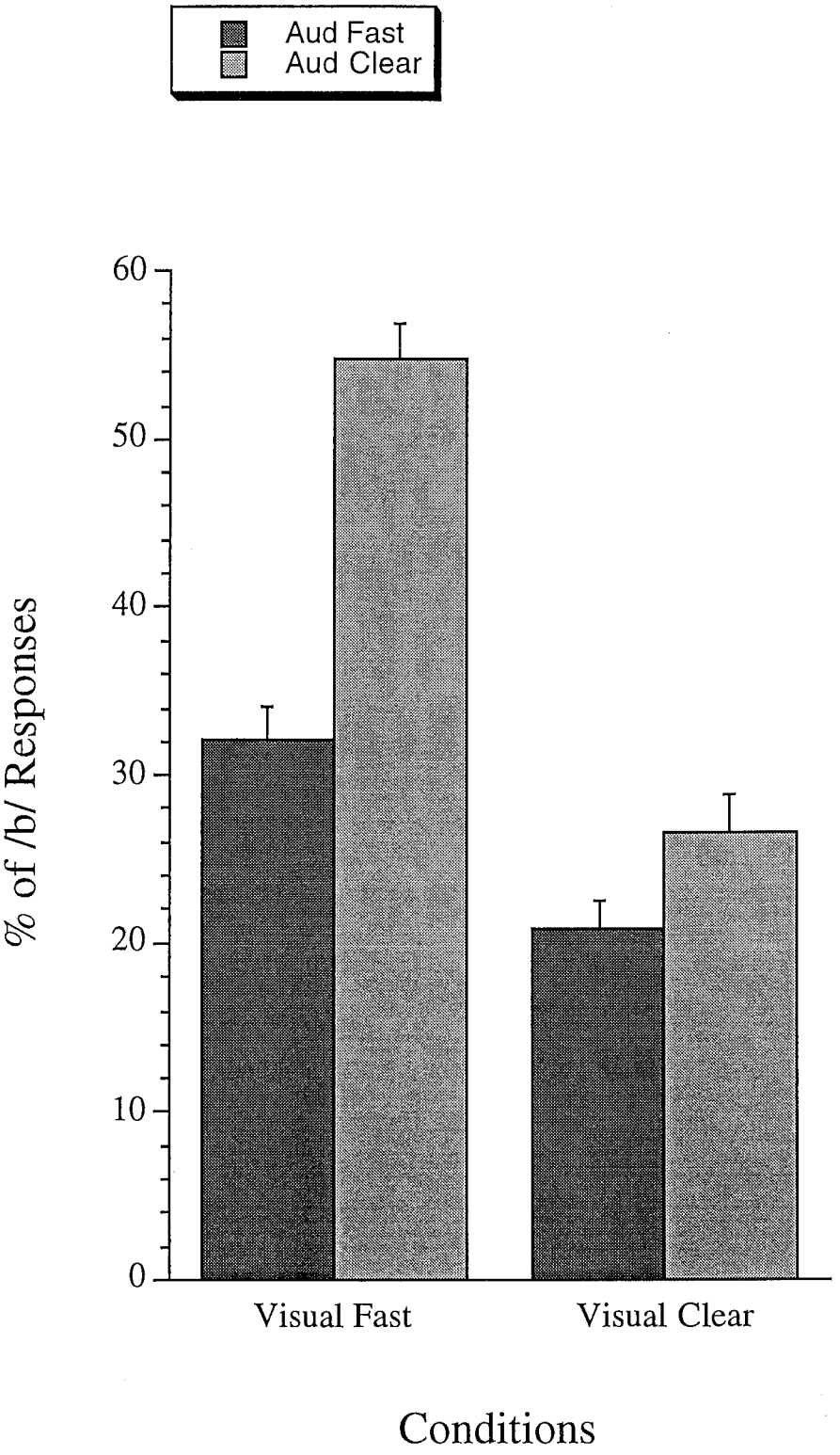

Figure 6. The percentage of /b/ responses as a function of the different visual and auditory speaking condition combinations. The error bars show the standard errors of the means.

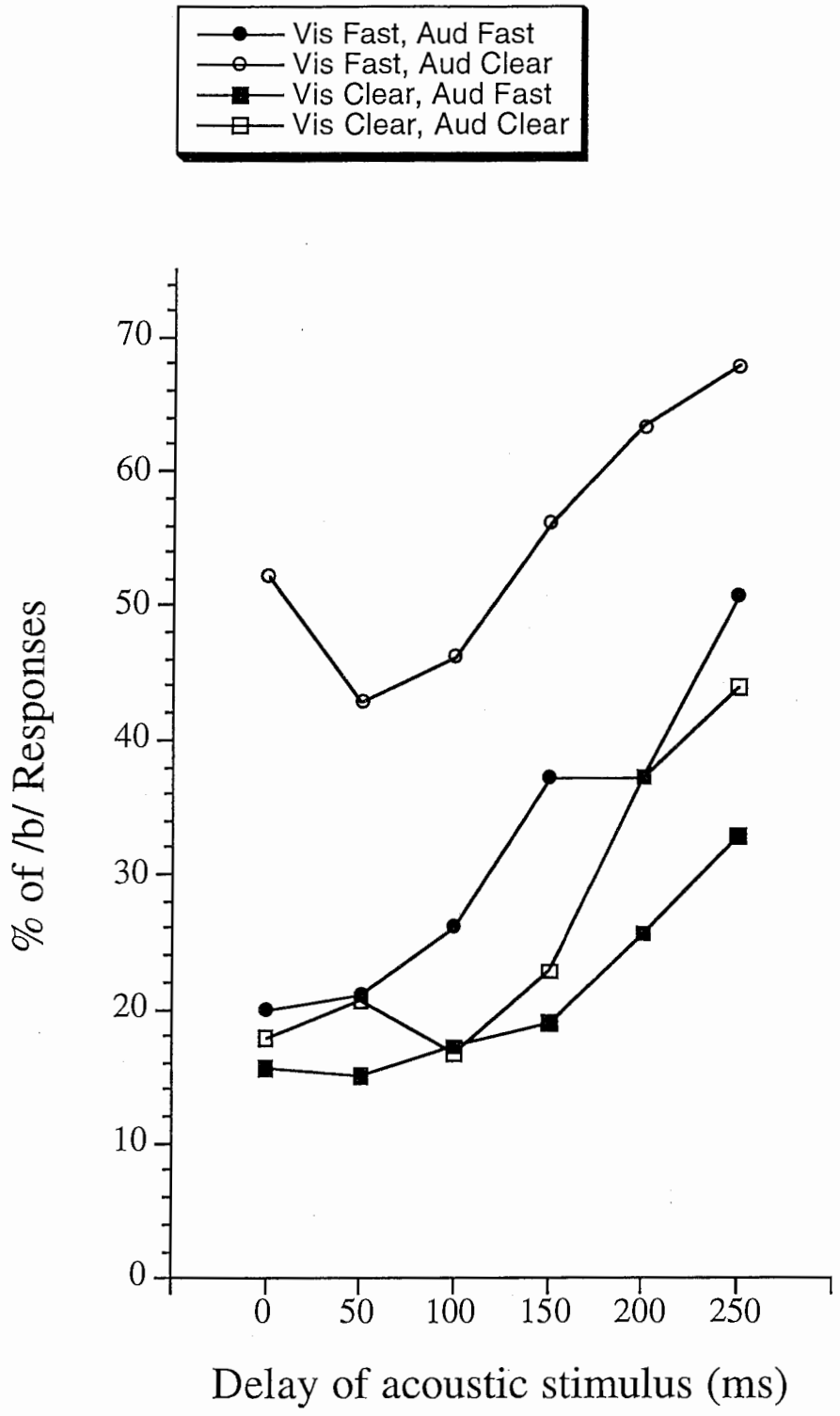

Figure 7. The percentage of /b/ responses as a function of the delay of the acoustic stimuli and the different visual and auditory speaking condition combinations.

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 ms. When the auditory signal led the visual stimuli, subjects showed less tolerance for the lack of synchrony. In all three experiments, the data indicated that the dynamic characteristics of articulation had a reliable effect on subjects' perception of the audiovisual stimuli. When the auditory and visual stimuli were produced under the same speaking conditions, subjects reported more McGurk effects.

The data are consistent with a body of work on the McGurk effect (e.g., Cohen, 1984; Green, 1994; Massaro & Cohen, 1993; Tillman et al., 1984) as well as research on synchrony in normal audiovisual productions (e.g., Dixon & Spitz, 1980; McGrath & Summerfield, 1984; 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 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 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 (e.g. Öhman, 1967) and syllables can vary in overall duration. This possibility is contradicted by the results of Experiment III. If syllable duration was a determining factor in the tolerance for audiovisual synchrony, we would have expected that the delay function for the visual fast/auditory fast 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.[4] 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 have an influence.

The differences associated with speaking condition reported in Experiments II and III suggest that speakers extract rate information from both modalities and that rate 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. While in the present data there seems to be 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 more influenced by the auditory rates of flickering bimodal stimuli than the visual rates. We cannot determine the source of this difference with the present data. As Green (1987) suggests, however, it may simply be that speaking rate and the type of rate measured by Welch et al. differ in terms of their auditory dominance.

In all three experiments the subjects showed sensitivity to the concordance of speaking dynamics between the two modalities. This finding has two implications. First, this results support the view that listeners use the time-varying properties of speech for perceptual grouping and phonetic perception (Remez and Rubin, 1991). 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 (e.g., Remez, Rubin, Berns, Pardo, & Lang, 1994). In sinewave speech, time-varying sinusoids track the

---

[4] It may be that the three speaking rates that we used here did not differ greatly in the duration of the critcal information about the moving vocal tract. Changes in rate are not uniformly distributed across syllables.

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 and, according to Remez and Rubin, this information is sufficient to specify that the sinewave stimuli are speechlike and it usually can lead to 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, & Demorest (1992) to account for subjects' ability to perceive haptic/acoustic and haptic/visual speech stimuli. This is not to say that dynamic information is the only cue that can aid audiovisual integration in speech. In our experiments, the stimuli in each modality are rich in information. As a result there are numerous clues to the identity of the tokens. However, it appears that one significant influence is how well the stimuli match in the information that they provide about the rate of change of the vocal tract.

Vocal tract movements are relatively slow (< 20 Hz) and thus the dynamic facial information has a limited frequency bandwidth. Recent evidence about the video frame rates necessary to support the visual perception of speech suggests that subjects need frame rates just fast enough to capture the motions of the face. Vitkovich & Barber (1994) suggest 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 findings is that they 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) states, rate-dependent information seem to be an obligatory part of speech processing, even in audiovisual perception.

Finally, we note that we still know relatively little about the time course of audiovisual information processing in speech perception. Smeele, Sittig, & van Heuvan (1994) have shown that subjects have access to visual information about place of articulation earlier than the auditory information. In part this was due to the fact that there is significant visual motion prior to the acoustic onset of the syllable. Information such as this on the timecourse of the processing of audiovisual perception will be important for understanding the nature of the integration of information form 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 rate of articulation within each of the modalities influences audiovisual perception. The results support the view that rate-dependent information is fundamental to phonetic perception.

Acknowledgments

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 and E. Vatikiotis-Bateson. We thank Lisa Coady for testing subjects in Experiment III and Jeff Jones for doing the video analysis in Experiment II.

References

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

BERNSTEIN, L.E., COULTER, D., O'CONNELL, EBERHART, S. & DEMOREST, M. (1992) Vibrotactile and haptic speech codes. Lecture presented at the Second International Conference on Tactile Aids, Hearing Aids, & Coclear Implants. Royal Institute of Technology, Stockholm, Swedan.

BERNSTEIN, L.E. & EBERHARDT, S. (1986). Audio-Visual Stimuli. Department of Electrical and Computer Engineering, John Hopkins University.

BRAIDA, D. L. (1991). Crossmodal integration in the Identification of Consonant Segments. *The Quarterly Journal of Experimental Psychology,* **43**, (3), 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 and Performance*, **17** (3), 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. Manuscript submitted to *Journal of the Academy of Rehabilatative Audiology.*

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

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

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

GREEN, K.P. (1994) Personal Communication.

GREEN, K. P. (1987). The perception of speaking rate using visual information from a talker's face. *Perception & Psychophysics,* **42** (6), 587-593.

GREEN, P. K., KUHL, K. P. & 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 meeting of the Acoustical Society of America, Honolulu.

GREEN, P. K. & KUHL, K. P. (1989). The role of visual information in the processing of place and manner features in speech perception. *Perception & Psychophysics*, **45** (1), 34-41.

GREEN, K. P., KUHL, P. K., MELTZOFF, A. N., STEVENS, E. R. (1991). Integrating speech information across talkers, gender, and sensory modality: Female

faces and male voices in the McGurk effect. *Perception & Psychophysics*, **50** (6), 524-536.

GREEN, K. & MILLER, J. (1985). On the role of visual rate information in phonetic perception. *Perception & Psychophysics*, **38** (3), 269-276.

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. *Proceedings of the International Conference on Spoken Language Processing*, 2047-2050.

LINDBLOM, B. (1990) Explaining phonetic variation: A sketch of the H and H theory. In W. Hardcastle and A. Marchal (eds.) *Speech Production and Speech Modeling*, Dordrecht: Kluwer.

MACDONALD, J. & MCGURK, H. (1978). Visual Influences on Speech Perception. *Perception and Psychophysics*, **24** (3), 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.* Erlbaum, Hillsdale.

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

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** (2), 678-684.

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. III Erlbaum, Hillsdale. 119-157.

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. *The 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 and Hearing Research*, **28**, 96-103.

PICK, H., WARREN, D., & HAY, J. (1969) Sensory conflict in judgements of spatial direction. *Perception and Psychophysics*, **6**, 203-205.

REMEZ, R. E. & RUBIN, P. E. (1991). Acoustic Shards, Perceptual Glue. To appear in J. Charles-Luce, P.A. Luce and J. R. Sawusch (Eds.), *Theories in Spoken Language: Perception, Production, and Development.* New Jersey: Ablex Press.

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. (1994) Temporal organization of bimodal speech information.*Proceedings of the International Conference on Spoken Language Processing*, 1431-1434.

SMEELE, P. M. T., SITTIG, A. C. & VAN HEUVEN, V.J. (1992) Intelligibility of audio-visually desychronised speech: asymmetrical effect of phoneme position. *Proceedings of the International Conference on Spoken Language Processing*. 65-68.

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.

TILLMANN, G. H. & POMPINO-MARSCHALL, U. P. (1984). Zum Einflub visuell dargeborener Sprechbewegungen auf die Wahrnehmung der akustisch kodierten Artikulation. Instituts fur Phonetik und Sprachliche Dommunikation der Universitat Munchen.

VATIKIOTIS-BATESON, E., EIGSTI, I. & YANO, S. (1994) Listener eye movement behavior during audiovisual speech perception. *Proceedings of the International Conference on Spoken Language Processing*, 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 and Hearing Research*, **37**, 1204-1210.

WARD, MEAGAN. (1992). *The effect of Auditory-Visual Dysynchrony on the Integration of Auditory and Visual Information in Speech Perception.* Unpublished thesis, Queens University, Kingston.

WELCH, R., DUTTONHURT, L., & WARREN, D. (1986). Contributions of vision and audition 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** (3), 638-667.