Audiovisual gating and the time course of speech perception

K. G. Munhall
Department of Psychology, Queen's University, Kingston, Ontario K7L 3N6, Canada

Y. Tohkura
ATR Human Information Processing Research Laboratories, Kyoto, Japan

(Received 4 March 1997; revised 27 February 1998; accepted 10 March 1998)

The time course of audiovisual information in speech perception was examined using a gating paradigm. VCVs that evoked the McGurk effect were gated visually and auditorily. The visual gating yielded a McGurk effect that increased in strength as a linear function of amount of visual stimulus presented. The acoustic gating revealed a more nonlinear function in which the VC information was considerably weaker than the CV portion of the VCV. The results suggest that the flow of cross-modal information is quite complex during audiovisual speech perception. © *1998 Acoustical Society of America.* [S0001-4966(98)05406-X]

PACS numbers: 43.71.Ma, 43.71.Es [WS]

INTRODUCTION

Visual and auditory information both contribute to the natural perception of speech [for reviews see Summerfield (1992) and Massaro (1987)]. In noisy conditions, for example, intelligibility increases if the speaker's face can be seen (Sumby and Pollack, 1954) and even under good listening conditions a visual stimulus can influence the identification of an auditory token (McGurk and MacDonald, 1976). In the present paper, we present evidence concerning the time course of both visual and auditory information for consonants. Specifically, we independently gated visual and auditory stimuli that elicit the McGurk effect to examine the temporal development of the audiovisual percept.

The McGurk effect is an audiovisual illusion in which one speech stimulus is presented auditorily (e.g., /bi/) and another is presented in synchrony visually (e.g., /gi/). Subjects frequently report hearing a third sound (e.g., /di/). This effect has been useful as a tool for the study of audiovisual integration in speech perception. While there have been numerous replications of McGurk and MacDonald's original finding (e.g., Green and Kuhl, 1989; Green *et al.*, 1988, 1991; MacDonald and McGurk, 1978; Manuel *et al.*, 1983; Massaro, 1987; Massaro and Cohen, 1983; Munhall *et al.*, 1996; Sekiyama and Tohkura, 1991; Summerfield and McGrath, 1984), a great deal is still unknown about the necessary and sufficient conditions required to produce audiovisual integration in speech. In this paper we explore how auditory and visual information for consonants unfolds over time.

The process of speech perception is necessarily extended in time because the information for individual sounds does not occur at any single instant (Liberman *et al.*, 1967). Both acoustic analyses and perceptual experiments have shown this extended span of perceptual information. From the production side this can be attributed to the overlap of speech gestures (e.g., Öhman, 1967; Fowler, 1977) and to the fact that individual speech gestures take a set amount of time. On the perceptual side, the temporal extent of perceptual processing is partly due to the nature of the information itself and partly due to the time course of the information processing. A number of studies have demonstrated that dynamic information extending over a syllable is important in vowel perception and normalization (Verbrugge *et al.*, 1976; Strange, 1989), auditory speech perception (Remez *et al.*, 1981), visual perception of speech (Rosenblum, 1994) and perhaps audiovisual integration (Munhall *et al.*, 1996).

There are many indications that the dynamic information from the visual and auditory modalities has a complicated timing structure. The visual information itself is asynchronous. The upper lip, lower lip and jaw, for example, have similar time functions but are shifted in phase with respect to each other (Gracco, 1988). In addition, the spans of visual and auditory information are not equal or coincident in time since there is frequently articulator motion before there is an acoustic consequence (Bell-Berti and Harris, 1981). This can be seen in preparatory adjustments for articulation or in the motion during the acoustic silence associated with the production of stop consonants (Löfqvist and Gracco, 1996); the tongue, jaw, and lips move continuously during oral closures. These acoustic/articulatory timing patterns often allow subjects to have access to visual information about place of articulation before the auditory information is available, e.g., before the silent interval for a stop is complete (Cathiard *et al.*, 1996; Green and Gerdman, 1995; Smeele, 1994).

In two experiments, we examined the relative timing of the information in the visual and auditory modalities. Further, we tested whether subjects can make use of the information available within a modality at any point in time. In the experiments we use a gating technique (Grosjean, 1980, 1996) in which increasing larger segments of the target stimulus are presented incrementally across trials. This is done by constructing a continuum in which the duration of the segment of the stimulus starting from the onset is increased in steps up to the full target stimulus duration. The interpretation of the results of gating experiments can be controversial. While some feel that the techniques show continuous perceptual uptake (e.g., Grosjean, 1996) others feel that gating is not an on-line task and that subjects are simply required to guess based on partial information (Cutler, 1995). Our experiments did not address this issue; rather, our aim was to examine the availability of information in the auditory and visual modalities at a point in time.

In our studies we separately gated the stimuli from the two modalities to study their individual contributions to the percept. In addition, we used stimuli in these gating studies that evoke the McGurk effect. The use of these contradictory stimuli allows us to test the separate contributions of the auditory and visual modalities. In the first experiment, the amount of visual information was varied while in the second experiment the amount of acoustic information was gated. In combination, the two studies allowed a comparison of the time course of audiovisual information available in McGurk stimuli. In particular, the experiments allowed us to compare the shape of gating functions obtained from the two modalities.

The shape of the gating function indicates the impact of equal temporal steps of information on perception and thus can serve as an index of the temporal development of information within a modality. Some theories of cross-modal perception (e.g., Welch and Warren, 1980) require a strong correlation between the information in the two modalities for audiovisual integration to occur, for such a theory to work, the temporal development of information should be similar for vision and audition. Testing this class of models requires that the cross-modal timing of information be specified in order to evaluate the temporal cohesion of the information across the two modalities. This includes determining the onset of information associated with a given segment within a modality as well as determining the time course of information within each modality. While we know that the visual information precedes acoustic information (e.g., Smeele, 1994), we know considerably less about the time course of information development in the two modalities during audiovisual perception. One reason for this is that, with a few notable exceptions (Cathiard *et al.*, 1996; Munhall *et al.*, 1996), the visual stimuli in audiovisual speech perception experiments are never characterized. It is common in the field to be provided only with information about the gender of the speaker and nothing about the dynamics of facial movement (Munhall and Vatikiotis-Bateson, 1998).

In the experiments presented here we provide an estimate of oral kinematics for each of the visual stimuli and we separately gate the auditory and visual tokens. The pattern of response or gating functions produced by subjects was used to estimate the available information within a modality. These gating functions for the two modalities establish the temporal conditions faced by any theory of cross-modal perception.

I. METHODS

A. Subjects

Seventy undergraduate students with self-reported normal hearing and normal or corrected to normal vision served as subjects in the two experiments. All subjects were native speakers of English. Twenty different subjects participated in the experimental conditions of each of the two experiments (40 subjects in total). An additional 20 subjects served in the control or baseline conditions of both experiments: Visual-only (experiment 1) and auditory-only (experiment 2) control conditions. Order of presentation of the two control conditions was balanced across subjects. Ten different subjects participated in a third baseline group in which no gating stimuli were presented. A between subject design was chosen for this experiment because pilot studies have shown that the magnitude of the McGurk effect is greatly influenced by subjects' experience with the auditory stimuli in auditory-only conditions. Thus different subjects were used for the auditory-only and audiovisual gating conditions. All subjects were naive to the purpose of the experiments and hadn't served in a McGurk experiment before.

B. Stimulus materials

In both experiments, natural productions of VCV stimuli by four different talkers were used. The stimuli consisted on a face saying /ægæ/ while the auditory stimuli were recordings of the same talkers saying /æbæ/. The visual stimuli were stored on a videodisc recorded at Queen's University. The auditory stimuli were digitized from the original sound tracts at 22-kHz sampling rate using a 12-bit A/D board (DataTranslation, DT2820).

C. Equipment

Subjects watched the displays on a 20-in. video monitor (Sony PVM 1910) and the videodiscs were played on a Pioneer (model LD-V8000) videodisc player. The acoustic signals were amplified, filtered with a 10-kHz cutoff 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.

D. Synchronization of stimuli

During the development of the experiments, the audio and visual stimuli were synchronized using the original sound track from the visual stimuli. We aligned the timing of the acoustic burst onset of the /g/ from the soundtrack of the /ægæ/ video with the burst onset of the acoustic stimulus, /b/. This timing relation was considered synchronous and the experimental software allowed this timing to be reliably reproduced (approximately 1-ms accuracy).

E. 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. Order of key labels was balanced across subjects. The first three labels stand for the stops /b/, /d/, /g/, and the final label stands for ''other.'' ''O'' was used when the subjects could not determine the consonant, when the subjects perceived a consonant other than one of the voiced stops, or when the did not hear a consonant at all. 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. In the video gating conditions the screen was blackened at the gated frame and remained black throughout the rest of the trial.

II. EXPERIMENT 1

Previous work on gating audiovisual stimuli (Cathiard *et al.*, 1996; Smeele, 1994) has indicated that the visual component of the stimuli contains strong information for place of articulation and that this information precedes the acoustic information in time. Smeele (1994) found that for CV stimuli the visual information was useful up to 150 ms prior to acoustic onset for the syllable. Similar results have been reported by Cathiard *et al.* (1996). In their study of V–V transitions, subjects were able to use gated visual information for vowel rounding up to 100 ms prior to the acoustic onset of the vowel.

In the first experiment, we explored the time course of the perception of the visual information in McGurk stimuli. The subjects viewed stimuli in which the visual intervocalic consonant was gated. In a typical trial, a subject would view a visual stimulus from stimulus onset to the gate location at which point the screen would go black for the rest of the trial. Different subjects served in the audiovisual and visual only conditions and gate durations were presented in random order.

A. Stimuli

The audio stimulus æbæ was always played in its entirety on each trial. As will be shown in experiment 2, the auditory stimuli were highly intelligible. The visual stimuli were gated in one frame steps from four frames before to four frames after the time of the acoustic burst. One frame equaled approximately 33 ms. In total, nine gate durations were used: From the beginning of the VCV to four, three, two, or one frame before the intervocalic burst, from the beginning of the VCV to (but not including) the intervocalic burst and from the beginning of the VCV to one, two, three, or four frames after the burst. The subjects were shown 160 trials in the experiment [4 talkers×4 repetitions×10 stimuli (9 gated stimuli+ungated stimulus)].

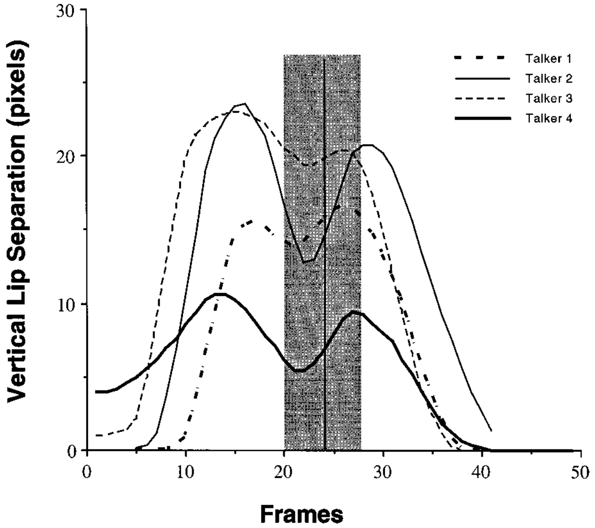

FIG. 1. Vertical separation of the lips in the visual stimuli produced by the four talkers. The vertical line indicates the frame containing the acoustic release burst. The shaded region indicates the range covered by the gating.

The gates were equal in duration for all talkers independent of the timing of their facial movements. Figure 1 shows the kinematics of oral opening for the four talkers and the location of the gates. In order to estimate the time course of the visual information, frame-by-frame measures of the vertical separation between the lips were recorded using a computerized image measurement system (Tiede, 1994).[1] For all talkers, the lips were in a closed position at the beginning of the trial, followed by a peak opening (approximately frame 15) for the first vowel. The oral aperture closes somewhat for the intervocalic /g/ and opens again for the second vowel (approximately frame 30). Finally, the oral aperture closes after the bisyllable. The vertical line indicates the frame during which oral release occurs. Thus the gated stimuli spanned the minimum aperture for the consonant and the movement toward the peak opening for the second vowel.

B. Results

Four responses were available to the subjects and the overall distribution of these responses can be seen in Fig. 2. The pattern of responses is influenced by the gating time with distinctly different patterns being observed for the four responses. In the audiovisual condition the ‘‘B’’ response decreases as the gate moves from four frames before to four frames after the intervocalic burst. ‘‘D’’ responses increase over the same interval while ‘‘G’’ and ‘‘O’’ response rates did not differ as a function of gate. The video-only control conditions are also plotted in each panel of Fig. 2. The ‘‘D’’ and ‘‘G’’ responses show an increase as a function of gate duration while the ‘‘B’’ responses are very infrequent and response rate is not influenced by gating time. The ‘‘O’’ decreases as a function of gate duration reflecting subjects’ inability to detect a consonant for the early gates.

ANOVAs were computed independently for each response. Main effects of modality (audiovisual versus video-only) were found for three of the response [$F(1,38) = 175.1$, 9.2, 99.8, $p<0.01$ for ‘‘B’’, ‘‘D’’, and ‘‘O’’, respectively] indicating that the percentage of ‘‘B’’ and ‘‘D’’ responses were higher in the audiovisual condition than the visual only. The percentage of ‘‘O’’ responses was higher in the visual-only than the audiovisual condition. Modality by gate time interactions were observed for three of the responses [$F(9,342) = 15.8$, 5.5, 21.2, $p<0.01$ for ‘‘B’’, ‘‘G’’, and ‘‘O’’, respectively; the gate time factor tests the nine gate times plus the ungated stimulus]. These interactions indicate that the size of the differences between audiovisual and visual-only conditions changed as a function of the gate time. For example, the rate of ‘‘O’’ responses is much greater in the visual-only condition at -4 frames than $+4$ frames (Fig. 2). Trend analyses were carried out to characterized the gating response functions in the audiovisual

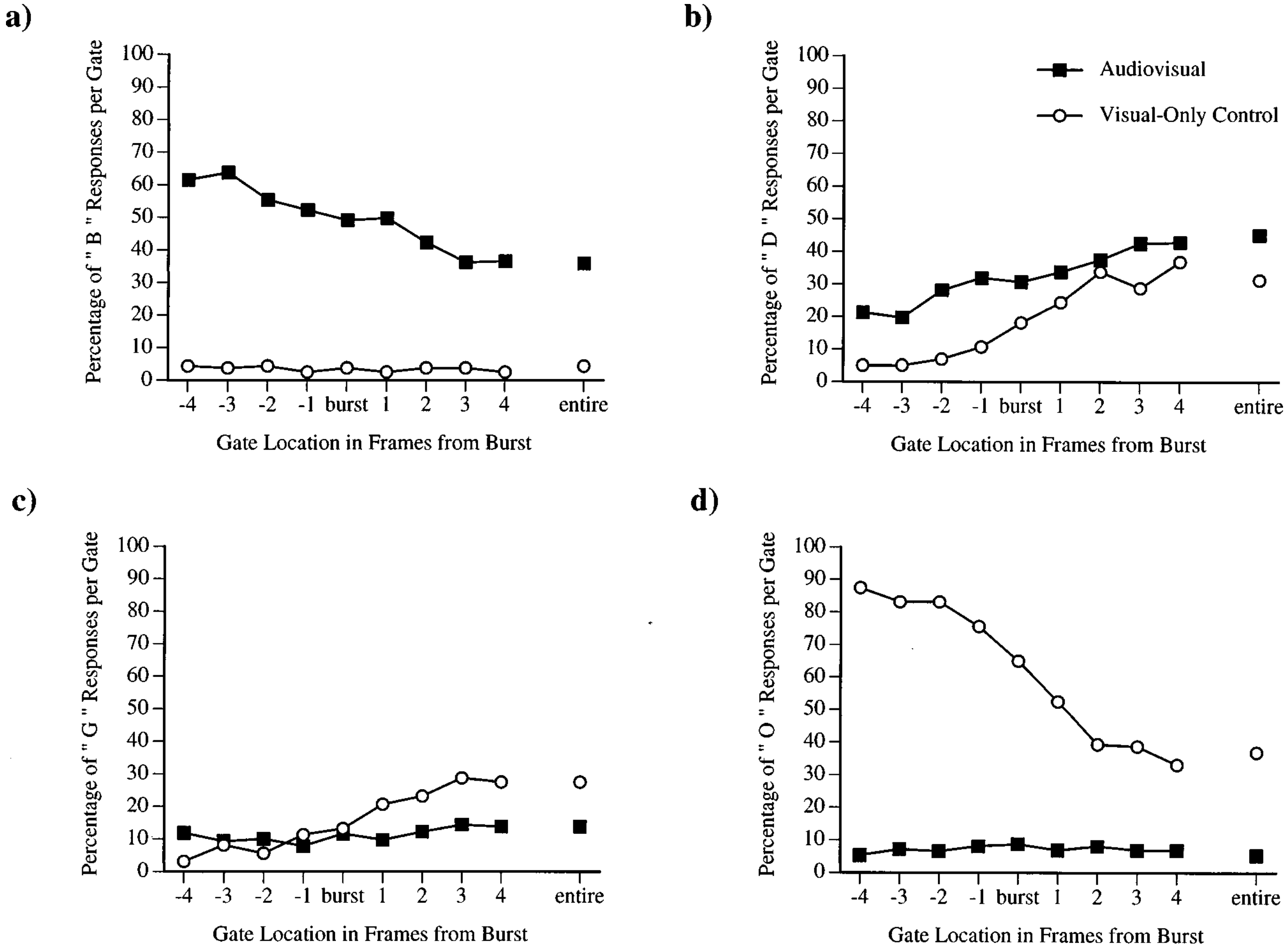

FIG. 2. Percentage of responses as a function of visual gate duration for (a) ''B'' responses, (b) ''D'' responses, (c) ''G'' responses, and (d) ''O'' or other responses. The boxes correspond to the audiovisual condition and the circles to the video-only condition.

condition. The ''B'' and ''D'' response options showed strong linear trends with no quadratic or cubic trends ($p < 0.0001$). The primarily linear functions for the visual conditions reflect the dynamical nature of the visual information. We are unable to determine from the gating technique whether the information is processed continuously (see Cutler, 1995) but it is clear that the information is incremental and is continuously available.

The stimulus set used in the experiment was produced by four different talkers whose tokens differed in the quality of visual and auditory information. Some talkers produce stimuli that yielded much stronger McGurk effects than others (Munhall *et al.*, 1996). By examining the gating functions for each talker we tested whether the average gating time function was observed for the individual talkers (i.e., for different levels of audiovisual integration) and whether the underlying pattern of visual intelligibility for the different talkers influenced the McGurk effect. In the visual only condition the number of reported ''B'' responses was consistently low for all talkers across gate durations [Fig. 3(a)]. The percentage of ''D'' responses [Fig. 3(b)], however, is somewhat higher for talker 1 and especially talker 4. The percentage of ''B'' and ''D'' responses in the audiovisual condition differ markedly for different talkers [Fig. 3(c), (d)]. Stimuli from talkers 2 and 4 evoked fewer ''B'' responses and more ''D'' responses than stimuli from talkers 1 and 3. However, the response functions from all 4 talkers were generally similar; they all showed a large linear trend with no obvious higher order trend.

The two talkers whose stimuli produced the largest McGurk effects (talker 2 and 4) had the largest movement associated with the intervocalic consonant. The decrease in aperture for the /g/ for talkers 1 and 3 was relatively modest compared to talkers 2 and 4. It should be noted that there was a substantial amount of articulator movement prior to the first gate, particularly for talkers 2 and 4. The use of this visual information might account for the substantial McGurk effects produced at the earliest gate. However, there does not seem to be any relationship between key events in the oral kinematics and the shape of the gating function. As can be seen by comparing Figs. 1 and 3, there is not a discontinuity in the response pattern at displacement or velocity peaks.

III. EXPERIMENT 2

In the second experiment, we examined the time course of audiovisual VCV perception by gating the auditory signal (Grosjean, 1980) at the same time points as used in experiment 1 before and after the intervocalic consonant burst. By presenting the auditory stimulus in increments we could assess the timing of the use of auditory information and the role the auditory information played in intermodal percep-

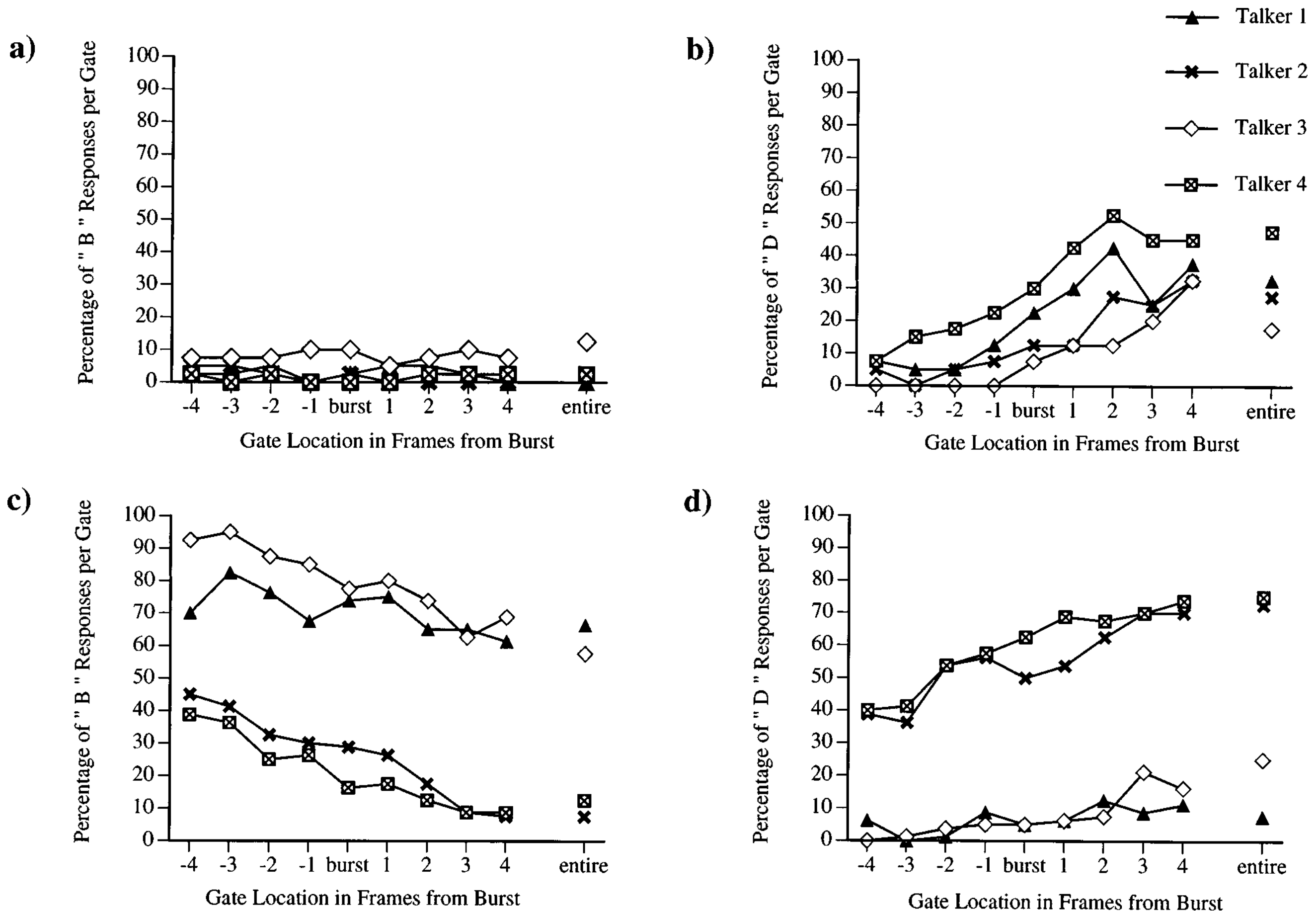

FIG. 3. Percentage of responses as a function of visual gate duration for the stimuli from the four talkers. (a) ''B'' responses in the visual only condition; (b) ''D'' responses in the visual only condition; (c) ''B'' responses in the audiovisual condition; (d) ''D'' responses in the audiovisual condition.

tion. Previous studies have shown that gating technique to be a robust (Cotton and Grosjean, 1984) and valid (Tyler and Wessels, 1985) tool for exploring the auditory perceptual information.

In this experiment we compared the perception of gated auditory stimuli in an auditory-only condition to gated auditory stimuli in an audiovisual condition to examine how the auditory contribution to the perceptual event grew over time. The different gate durations were presented in random order within a modality condition. Different individuals served as subjects in the auditory-only and audiovisual conditions.

A. Stimuli

The video stimuli were played in their entirety on each trial. The auditory stimuli were presented in nine gate durations: From the beginning of the VCV to 120, 90, 60, and 30 ms before the intervocalic burst, from the beginning of the VCV to (but not including) the intervocalic burst and from the beginning of the VCV to 30, 60, 90, and 120 ms after the burst. These gate intervals corresponded approximately to the timing of gates of one frame used in the visual gating in experiment 1.

B. Results

Four responses were available to the subjects and the overall distribution of these responses can be seen in Fig. 4. The pattern of responses is influenced by the gating time with distinctly different patterns being observed for the four responses. In the audiovisual condition the ''O'' response decreased as the gate moved from 120 ms before to 120 ms after the intervocalic burst. The other three responses (''B, D, G'') increased over the same interval with ''B'' and ''D'' increasing more than ''G''. All four responses showed major changes in response rate between the gates at the intervocalic burst and the gate 30 ms later. The auditory-only control conditions are also plotted in Fig. 4. The ''B'' and ''O'' responses showed a similar pattern, while the ''D'' and ''G'' responses were very infrequent and response rate was not influenced by gating time.

ANOVAs were computed independently for each response. Main effects of modality (audiovisual versus audio-only) were found for all four responses [$F(1,38) = 168.1$, 70.2, 18.0, 11.3, $p<0.01$ for ''B'', ''D'', ''G'', and ''O'', respectively]. Modality by gate time interactions were observed for all four responses [$F(9,342) = 14.2$, 45.8, 4.1, 6.6, $p<0.01$ for ''B'', ''D'', ''G'', and ''O'', respectively].

Trend analyses were carried out to characterize the gating response functions in the audiovisual condition. Unlike the pattern observed in experiment 1 for visual information, the pattern was significantly nonlinear. Three of four response options (with ''G'' being the exception) showed strong linear and cubic trends ($p<0.0001$). In large part, this

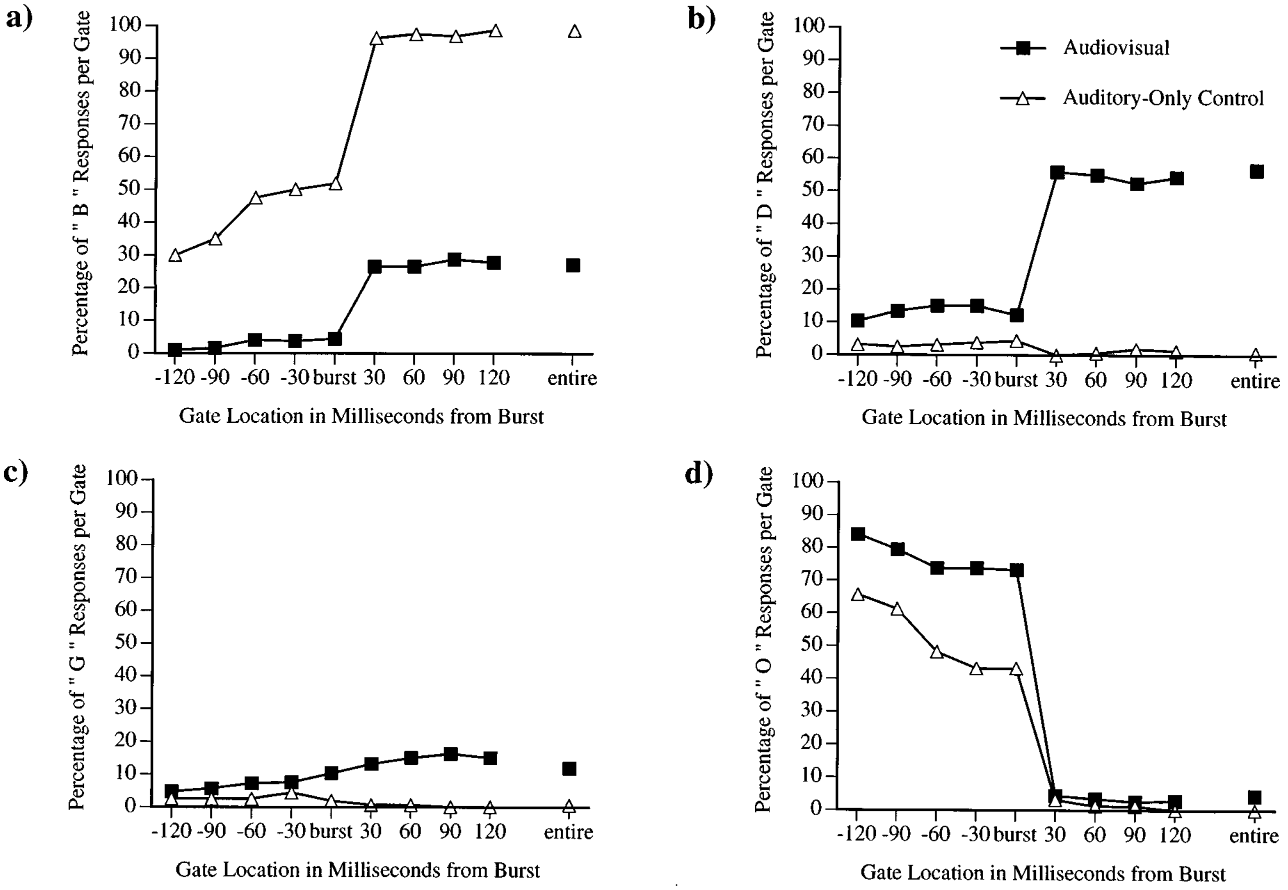

FIG. 4. Percentage of responses as a function of auditory gate duration for (a) ''B'' responses, (b) ''D'' responses, (c) ''G'' responses, and (d) ''O'' or other responses. The boxes correspond to the audiovisual condition and the triangles to the auditory-only condition.

pattern was determined by the particular gate locations used with these stimuli. The gating manipulation in this experiment was carried out using gate steps of equal duration for all four talkers' stimuli, chosen to roughly correspond with the timing of visual gates used in experiment 1 (i.e., approximately 1 video frame). Table I shows the durations of the acoustic intervals in the four talkers' /æbæ/ productions. Two things should be noted about the data in this table. First, the stimuli were not all equal in duration and this means that the gates were not always in equivalent positions. Second, the gates preceding the burst occurred mostly during oral closure for the /b/. Voicing was apparent during all of the talkers' closures but the subjects did not make use of the closure information until the duration of the interval was signaled. Thus, the acoustic stimuli seem to be divided into VC and CV portions by the gating. The one exception to this pattern was talker 3, whose closure was quite short and thus the gates preceding the burst occurred mostly during the first vowel. Interestingly, the percentage of /b/s reported for the auditory only condition of talker 3 showed a distinctly different pattern than the other three talkers [Fig. 5(a)]. However, all talkers showed a similar nonlinear pattern in the audiovisual condition [Fig. 5(b) and (c)]. While the auditory information may contain dynamical cues, it appears to be punctuated by discontinuities that have different degrees of perceptual impact. The visual stimuli in this task provided continuous information over time while auditory perception was marked by instants or time windows of greater or lesser information (cf. Blumstein and Stevens, 1980).

TABLE I. Durations of the acoustic stimuli used in experiments 1 and 2.

Talker	Vowel 1	Closure	VOT	Vowel 2
1	139.8	119.5	11.0	233.3
2	215.8	139.6	7.0	286.9
3	159.1	69.6	4.8	246.9
4	244.9	115.8	5.7	323.9

The pattern of response in the auditory-only condition was consistent with previous findings on the perception of intervocalic consonants [Dorman *et al.*, 1979; Householder, 1956 (cited in Byrd, 1992); Streeter and Nigro, 1979]. While information for the consonant may be present in both VC and the CV portions of the bisyllable, information in the CV portion appears to be much more salient. Subjects did not report a large number of ''B'' responses in the auditory-only condition until they heard the release burst (gate 30 ms after the burst). The discontinuity in the response functions reflect the differential strength of the VC versus CV cues.

The pattern of responses in the audiovisual condition largely mirrored the pattern observed for the auditory only. For example, ''B'' and ''O'' response functions were similar in shape for the audiovisual and auditory-only conditions. The largest difference between the two conditions and the largest interaction effect was observed for ''D'' responses.

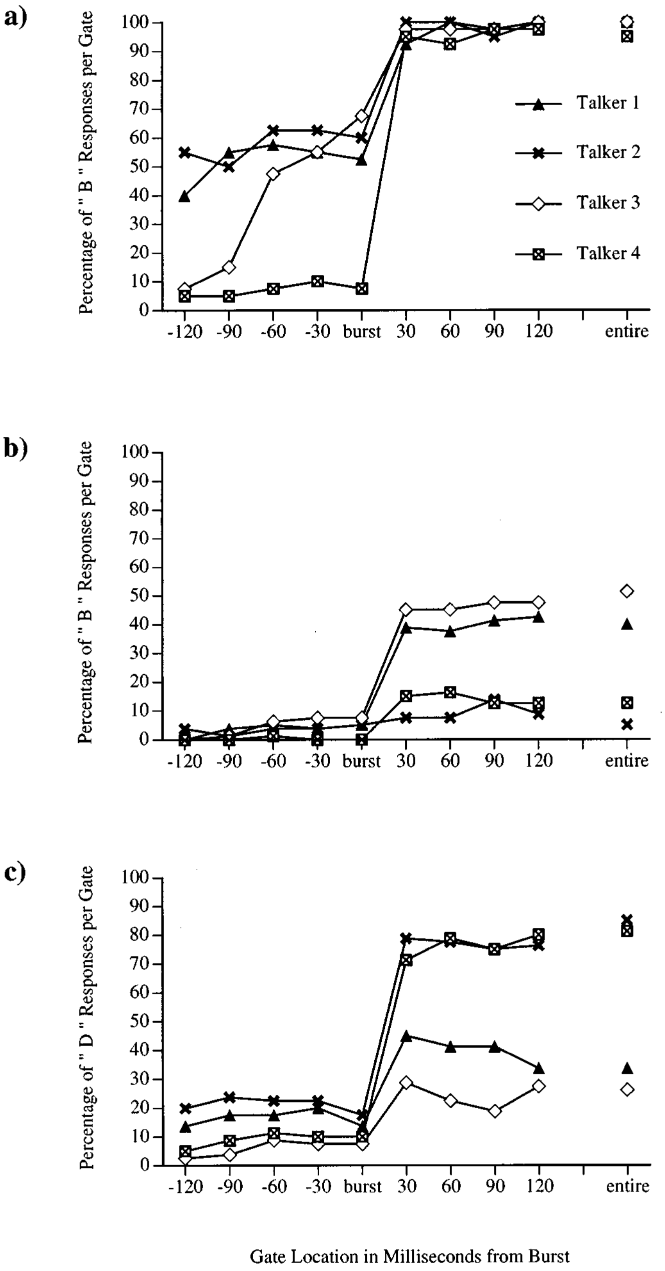

FIG. 5. Percentage of responses as a function of auditory gate duration for the stimuli from the four talkers. (a) ''B'' responses in the auditory only condition; (b) ''B'' responses in the audiovisual condition; (c) ''D'' responses in the audiovisual condition.

This is interesting in that it suggests that until there is sufficient auditory information to support a strong ''B'' response (or at least to support the perception of a consonant) the McGurk effect cannot be observed.

As in experiment 1, we examined the patterns of responses for stimuli produced by the different talkers. Figure 5 shows the percentage of /b/ and /d/ responses for the audiovisual condition and the percentage of /b/ responses for the auditory-only condition. The /b/ responses for the auditory-only condition [Fig. 5(a)] reveal that accuracy for stimuli from all talkers reached ceiling at the first gate beyond the intervocalic burst. While perception of the /b/ varied for different talkers before the burst, there was little difference among the four talkers' asymptotic response levels following the burst. The differences in perception of /b/ in the auditory-only condition before the burst do not directly account for the talker differences in elicitation of the McGurk effect. Talker 4's stimuli produced by far the lowest perception of /b/ while the stimuli for talker 2 produced the highest perception of /b/. The stimuli for these two talkers showed the strongest McGurk effect overall [Fig. 5(c)].

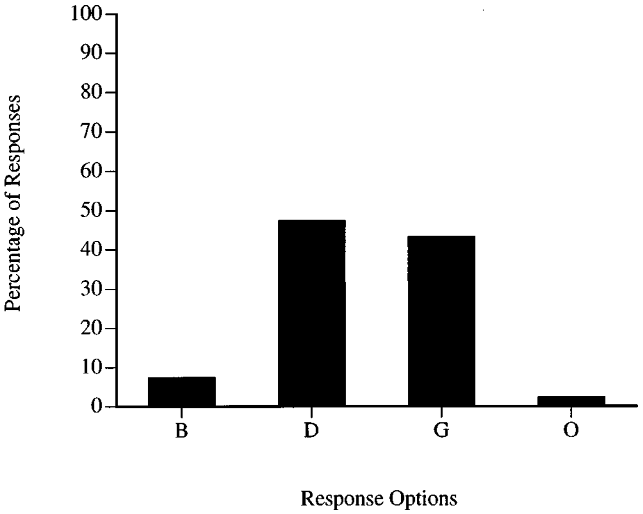

FIG. 6. Percentage of ''B,'' ''D,'' ''G,'' and ''O'' responses in the audiovisual control condition.

In Fig. 4, the percentage of ''B'' responses given when the ungated audiovisual stimulus was played is shown. Both the value in this experiment and the value reported for the same condition in experiment 1 are higher than we usually observe for stimuli of this kind. To verify this we tested an additional 10 subjects who were only shown the complete audiovisual McGurk stimuli (i.e., no gated stimuli were presented). Figure 6 shows the average responses for these subjects. The percentage of ''B''s reported by these subjects (7.3%), for example, is much lower than reported in experiments 1 (35.9%) and 2 (27.2%). It would appear either that the gated stimuli induced a bias or that sampling differences produced this pattern of results (see below).

IV. GENERAL DISCUSSION

The data from the two experiments show distinct differences between the visual and auditory signals and how they are processed. The auditory information for the intervocalic consonant appears to be much stronger in the CV portion of the bisyllable than in the VC portion. As a result auditory consonant identification was quite poor until after the intervocalic release burst was heard. The perception of the visual information, on the other hand, was more continuous. The percentage of /b/s reported in the audiovisual condition of the visual gating experiment (exp. 1) changed as a linear function of the amount of the visual stimulus presented. These patterns of results were observed for three groups of subjects and were observed for stimuli produced by four different talkers.

This pattern suggests that visual information unfolds smoothly with the dynamics of articulation but that the acoustic information does not always directly reflect these smooth kinematics. Rather, the acoustics are punctuated by instants of greater information value. This can occur for a number of reasons. Some acoustic instants may be highly salient because the cues at a point in time are robust. For example, the burst of a stop which extends over a small temporal span is very informative. A second source of nonlinearity in the acoustic information is the perceptual use of temporal intervals. A stretch of silence has little or no information value until it is completed. Only at the end of the interval can the duration indicate the manner and style of articulation.

Previous research on the time course of audiovisual speech perception has indicated that the visual cues for place of articulation can precede the acoustic information by more than 100 ms (e.g., Smeele, 1994). This research, in conjunction with the present findings, suggests that the flow of cross-modal information is quite complex. Information from the visual and auditory modalities is not synchronous and does not unfold continuously at the same rate. Thus it seems unlikely that the perceptual system uses temporal coincidence or any tight cross-modal timing as the basis of audiovisual integration. Studies in which the synchrony of the auditory and visual signals was directly manipulated support this conclusion (e.g., Abry *et al.*, 1996; Massaro and Cohen, 1993; Munhall *et al.*, 1996; Tillman *et al.*, 1984). For example, Munhall *et al.* found that the acoustic signal could be delayed up to 180 ms from normal audiovisual timing without a significant decrease in the strength of the McGurk effect.

While the perceptual system is quite tolerant of desynchrony of audiovisual information, there are limitations. This is particularly true when the acoustic information is advanced in time relative to the visual stimulus. A number of experiments have shown that sound-lead is not tolerated as well as sound-lag (e.g., Abry *et al.*, 1996; Dixon and Spitz, 1980; Munhall *et al.*, 1996; Smeele *et al.*, 1992; Tillman *et al.*, 1984; but cf. Gerdeman, 1994). Abry *et al.* suggest that the tolerance for advancing the sound is limited by the natural audiovisual timing relationship. Articulatory movements naturally precede the acoustics and the sound can be advanced only until the point at which the acoustic information would precede the gestures. This cross-modal timing boundary might suggest a limited perceptual span over which the information from the two modalities must be linked and thus some temporal dependence in cross-modal perception. In our view a more likely explanation is that the limitation is not time *per se* but rather the information processing of the auditory information; visual information might not influence perceptual categorization if the auditory information has reached a criterion threshold. The more natural temporal precedence of visual information may have led to a priming role for the visual signal in normal audiovisual perception. Sams *et al.* (1991) have also concluded that the visual stimuli may prime the auditory stimuli based on their magnetoencephalographic study of the McGurk effect. While auditory segmental context effects (in which later occurring information influences the category judgment for a preceding segment) are numerous (e.g., Mann and Repp, 1980), visual information may not produce similar effects on auditory perception.

Two studies may provide counterevidence to the suggestion that visual information that follows auditory information in time does not produce context effects on auditory perception. Green and Gerdeman (1995) have shown that the magnitude of the McGurk effect is influenced by whether the following visual vowel matches the auditory vowel. Green and Miller (1985) showed that the voicing boundary in an auditory /bi–pi/ continuum was influenced by the visual speaking rate for the bilabial. One of the primary differences between speaking rates was the duration of the visual vowels. However, in both of these studies that were also visual cues for speaking rate or vowel context that preceded the auditory information for the consonant. The visual information for different speaking rates is based in the dynamics of articulation which are apparent throughout the visual syllable. For example, velocities of speech movement can change with speaking rate and peak velocities occur quite early within a gesture. Similarly, the articulation of consonants is influenced quite early by following vowels and therefore, the coarticulation effect reported by Green and Gerdeman (1995) may not be due to visual information that follows the consonant. Thus it is possible to interpret even these studies as instances of visual priming of the auditory judgment.

It should be noted that the pattern of results for the visual gating experiment indicates that the visual information is continuously available and incrementally useful to the perceiver but it does not speak to the issue of the continuity of uptake of the visual information. The gating paradigm requires subjects to guess on the basis of the available information for each gate duration. The observed results do not indicate that the time course of the normal perceptual decision process mirrors the gating function.

The kinematics of the lip movements shown in Fig. 1 represent only a part of the complex visual stimulus provided by the moving face [see Munhall and Vatikiotis-Bateson (1998) for a review of facial dynamics during speech]. It is clear even from this crude measure, however, that the face is continuously varying during speech production. This time-varying visual information signals the rate of speaking (Green, 1987; Green and Miller, 1985), as well as indicating segmental structure and segmental categories (Summerfield, 1987). A number of years ago Remez *et al.* (1981) showed that time-varying acoustic information is sufficient to signal the phonetic structure of an utterance. In a previous paper (Munhall *et al.*, 1996) we suggested that this dynamic information may play a crucial role in audiovisual integration. The complexity of the timing data in the present studies raises the possibility that the dynamic information is extracted separately for each modality before the information from vision and audition is merged. Recently, Green *et al.* (1991) and Massaro (1987) have argued similarly that auditory and visual information is processed to some degree before integration takes place. While Green *et al.* argue only that talker characteristics may be processed, Massaro proposes that the auditory and visual information are completely processed prior to integration.

One of the intriguing observations in these experiments is that subjects who were presented only with complete audiovisual VCVs (Fig. 6) showed a stronger McGurk effect than the experimental subjects in the two studies (''entire'' in Figs. 2 and 4). Two possible explanations are available for this finding. First, chance sampling differences may account for the discrepancy between the subjects. Second, the pattern of results may be caused by contextual biasing. In the experimental group that perceived the highest number of /b/s across all gating conditions (subjects in exp. 1), the weakest McGurk effect was found for the ungated stimuli. In the group that perceived the next highest number of /b/s across all gating conditions (exp. 2), an intermediate McGurk effect was observed for the ungated stimuli. Finally, the McGurk-only group (i.e., no gating) that had the least experience with auditory /b/ perceptions overall showed the strongest McGurk effect. While these trends are not conclusive, they are consistent with the idea that exposure to /b/ exemplars may alter or bias responses in the McGurk effect (cf. Case *et al.*, 1995).

ACKNOWLEDGMENTS

This research was funded by NIH Grant No. DC-00594 from the National Institute of Deafness and other Communicative Disorders and NSERC. The authors wish to thank Kerry Green and Eric Vatikiotis-Bateson for helpful comments on an earlier version of this paper.

[1] Use of this measure is a simplification. No single point on the face captures all of the visual information available to a perceiver. A number of recent perceptual studies have demonstrated that the motion of the full face is used in speechreading (Smeele *et al.*, 1995; Guiard-Marigny *et al.*, 1995). In these studies, subjects performed better when more of the facial surface than the lips was visible. Statistical examination of the motion of various regions of the face is consistent with these findings (Vatikiotis-Bateson *et al.*, 1996). When the 3-D motion of markers on the face was used to estimate the rms amplitude of the speech acoustics, the best estimate was achieved using all of the markers including positions at a distance from the lips.

Abry, C., Lallouache, M. T., and Cathiard, M. A. (**1996**). ''How can coarticulation models account for speech sensitivity to audio-visual desynchronization?,'' in *Speechreading by Humans and Machines*, edited by D. Stork and M. E. Hennecke (Springer-Verlag, Berlin).

Bell-Berti, F., and Harris, K. S. (**1981**). ''A temporal model of speech production,'' Phonetica **38**, 9–20.

Blumstein, S. E., and Stevens, K. N. (**1980**). ''Perceptual invariance and onset spectra for stop consonants in different vowel environments,'' J. Acoust. Soc. Am. **67**, 648–662.

Byrd, D. (**1992**). ''Perception of assimilation in consonant clusters: A gestural model,'' Phonetica **49**, 1–24.

Case, P., Tuller, B., Ding, M., and Kelso, J. A. S. (**1995**). ''Evaluation of a dynamical model of speech perception,'' Percept. Psychophys. **57**, 977–988.

Cathiard, M. A., Lallouache, M. T., and Abry, C. (**1996**). ''Does movement on the lips mean movment in the mind?,'' in *Speechreading by Humans and Machines*, edited by D. Stork and M. E. Hennecke (Springer-Verlag, Berlin).

Cotton, S., and Grosjean, F. (**1984**). ''The gating paradigm: A comparison of successive and individual presentation formats,'' Percept. Psychophys. **35**, 41–48.

Cutler, A. (**1995**). ''Spoken word recognition and production,'' in *Speech, Language, and Communication*, edited by J. L. Miller and P. D. Eimas (Academic, San Diego).

Dixon, N., and Spitz, L. (**1980**). ''The detection of audiovisual desynchrony,'' Perception **9**, 719–721.

Dorman, M. F., Raphael, L. J., and Liberman, A. M. (**1979**). ''Some experiments on the sound of silence in phonetic perception,'' J. Acoust. Soc. Am. **65**, 1518–1532.

Fowler, C. A. (**1977**). *Timing Control in Speech Production* (Indiana University, University Club, Bloomington).

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

Gracco, V. L. (**1988**). ''Timing factors in the coordination of speech movements,'' J. Neurosci. **8**, 4628–4634.

Green, K. P. (**1987**). ''The perception of speaking rate using visual information from a talker's face,'' Percept. Psychophys. **42**, 587–593.

Green, K., and Miller, J. (**1985**). ''On the role of visual rate information in phonetic perception,'' Percept. Psychophys. **38**, 269–276.

Green, K. P., and Gerdeman, A. (**1995**). ''Cross-modal discrepancies in coarticulation and the integration of speech information: The McGurk effect with mismatched vowels,'' J. Exp. Psychol. **21**, 1409–1426.

Green, K. P., and Kuhl, K. P. (**1989**). ''The role of visual information in the processing of place and manner features in speech perception,'' Percept. Psychophys. **45**, 34–41.

Green, K. P., Kuhl, K. P., and Meltzoff, N. A. (**1988**). ''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, K. P., Kuhl, K. P., Meltzoff, A. N., and Stevens, E. R. (**1991**). ''Integrating speech information across talkers, gender, and sensory modality: Female faces and male voices in the McGurk effect,'' Percept. Psychophys. **50**, 524–536.

Grosjean, F. (**1980**). ''Spoken word recognition processes and the gating paradigm,'' Percept. Psychophys. **28**, 267–283.

Grosjean, F. (**1996**). ''Gating,'' Language and Cognitive Processes **11**, 597–604.

Guiard-Marigny, T., Ostry, D. J., and Benoit, C. (**1995**). ''Speech intelligibility of synthetic lips and jaw,'' *Proceedings of the 13th International Congress of Phonetic Sciences* (Stockholm, Sweden), Vol. 3, pp. 222–225.

Liberman, A., Cooper, F., Shankweiler, D., and Studdert-Kennedy, M. (**1967**). ''Perception of the speech code,'' Psychol. Rev. **74**, 431–461.

Löfqvist, A., and Gracco, V. (**1996**). ''Labial kinematics in stop consonant production,'' J. Acoust. Soc. Am. **99**, 2472 (A).

Mann, V., and Repp, B. (**1980**). ''Influence of vocalic context on perception of the [ʃ]–[s] distinction,'' Percept. Psychophys. **28**, 213–228.

Manuel, S. Y., Repp, B., Studdert-Kennedy, M., and Liberman, A. (**1983**). ''Exploring the McGurk effect,'' J. Acoust. Soc. Am. Suppl. 1 **74**, S66.

MacDonald, J., and McGurk, H. (**1978**). ''Visual influences on speech perception,'' Percept. Psychophys. **24**, 253–257.

Massaro, D. W. (**1987**). *Speech Perception by Ear and Eye* (Erlbaum, Hillsdale, NJ).

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

McGurk, H., and MacDonald, J. (**1976**). ''Hearing lips and seeing speech,'' Nature **264**, 746–748.

Munhall, K. G., Gribble, P., Sacco, L., and Ward, M. (**1996**). ''Temporal constraints on the McGurk effect,'' Percept. Psychophys. **58**, 351–362.

Munhall, K. G., and Vatikiotis-Bateson, E. (**1998**). ''The moving face during speech communication,'' in *Hearing by Eye, Part 2: The Psychology of Speechreading and Audiovisual Speech*, edited by R. Campbell, B. Dodd, and D. Burnham (Taylor and Francis, Psychology Press, London).

Nusbaum, H. C., and Morin, T. M. (**1992**). ''Paying attention to differences among talkers,'' in *Speech Perception, Production and Linguistic Structure*, edited by Y. Tohkura, E. Vatikiotis-Bateson, and Y. Sagisaka (Ohmsha, Tokyo).

Öhman, S. (**1967**). ''Numerical model of coarticulation,'' J. Acoust. Soc. Am. **41**, 310–320.

Remez, R., Rubin, P., Pisoni, D., and Carrell, T. (**1981**). ''Speech perception without traditional cues,'' Science **212**, 947–950.

Rosenblum, L. (**1994**). ''How special is audiovisual speech integration?,'' Current Psychology of Cognition **13**, 110–116.

Sams, M., Aulanko, R., Hämälänin, M., Hari, R., Lounasmaa, O., Lu, S., and Simola, J. (**1991**). ''Seeing speech: Visual information from lip movements modifies activity in the human auditory cortex,'' Neurosci. Lett. **127**, 141–145.

Sekiyama, K., and Tohkura, Y. (**1991**). ''McGurk effect in non-English listeners: Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility,'' J. Acoust. Soc. Am. **90**, 1797–1805.

Smeele, P. M. T. (**1994**). ''Perceiving speech: Integrating auditory and visual speech,'' unpublished doctoral dissertation, Delft University of Technology.

Smeele, P. M. T., Sittig, A. C., and Van Heuven, V. J. (**1992**). ''Intelligibility of audio-visually desynchronized speech: asymmetrical effect of phoneme position,'' *Proceedings of the International Conference on Spoken Language Processing*, pp. 65–68.

Smeele, P., Hahnlen, L., Stevens, E., Kuhl, P., and Meltzoff, A. (**1995**). ''Investigating the role of specific facial information in audiovisual speech perception,'' J. Acoust. Soc. Am. **98**, 2569(A).

Strange, W. (**1989**). ''Evolving theories of vowel perception,'' J. Acoust. Soc. Am. **85**, 2081–2087.

Streeter, L., and Nigro, G. (**1979**). ''The role of medial consonant transitions in word perception,'' J. Acoust. Soc. Am. **65**, 1533–1541.

Sumby, W. H., and Pollack, I. (**1954**). ''Visual contribution to speech intelligibility in noise,'' J. Acoust. Soc. Am. **26**, 212–215.

Summerfield, Q. (**1987**). ''Some preliminaries to a comprehensive account of audio-visual speech perception,'' in *Hearing by Eye: The Psychology of Lip-Reading*, edited by B. Dodd and R. Campbell (Erlbaum, London), pp. 3–51.

Summerfield, Q. (**1992**). ''Lipreading and audio-visual speech perception,'' Philos. Trans. R. Soc. London, Ser. B **335**, 71–78.

Summerfield, Q., and McGrath, M. (**1984**). ''Detection and resolution of audio-visual incompatibility in the perception of vowels,'' Q. J. Exp. Psychol. **36A**, 51–74.

Tiede, M. (**1994**). Vidiot (Macintosh video analysis software), ATR Laboratories, Kyoto.

Tillman, H. G., Pomino-Marschall, B., and Porzig, H. (**1984**). ''Zum Einfluß visuell dargeborener Sprachbewegungen auf die Wahrnehumung der akustisch kodieten Artikulation,'' Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation der Universität München **19**, 318–338.

Tyler, L. K., and Wessels, J. (**1985**). ''Is gating an on-line task? Evidence from naming latency data,'' Percept. Psychophys. **35**, 409–420.

Vatikiotis-Bateson, E., Munhall, K. G., Kasahara, Y., Garcia, F., and Yehia, H. (**1996**). ''Characterizing audiovisual information during speech,'' *Proceedings of the Fourth International Conference on Spoken Language Processing*, ICSLP-96, pp. 1485–1488.

Verbrugge, R., Strange, W., Shankweiler, D., and Edman, T. (**1976**). ''What information enables a listener to map a talker's vowel space?,'' J. Acoust. Soc. Am. **60**, 198–212.

Welch, R. B., and Warren, D. H. (**1980**). ''Immediate perceptual response to intersensory discrepancy,'' Psychol. Bull. **88**, 638–667.