1408

Perception & Psychophysics
2005, 67 (5), 759-769

Use of visual information in speech perception: Evidence for a visual rate effect both with and without a McGurk effect

LAWRENCE BRANCAZIO
Southern Connecticut State University, New Haven, Connecticut and Haskins Laboratories, New Haven, Connecticut

and

JOANNE L. MILLER
Northeastern University, Boston, Massachusetts

The McGurk effect, where an incongruent visual syllable influences identification of an auditory syllable, does not always occur, suggesting that perceivers sometimes fail to use relevant visual phonetic information. We tested whether another visual phonetic effect, which involves the influence of visual speaking rate on perceived voicing (Green & Miller, 1985), would occur in instances when the McGurk effect does not. In Experiment 1, we established this visual rate effect using auditory and visual stimuli matching in place of articulation, finding a shift in the voicing boundary along an auditory voice-onset-time continuum with fast versus slow visual speech tokens. In Experiment 2, we used auditory and visual stimuli differing in place of articulation and found a shift in the voicing boundary due to visual rate when the McGurk effect occurred and, more critically, when it did not. The latter finding indicates that phonetically relevant visual information is used in speech perception even when the McGurk effect does not occur, suggesting that the incidence of the McGurk effect underestimates the extent of audiovisual integration.

A well-established finding is that the perception of speech can be highly influenced by vision. The evidence for visual contributions to speech perception has largely come from two lines of research. The first line involves studies in which auditory speech stimuli are presented in noise, either alone or with a simultaneous visual presentation of the talker's articulating face; the intelligibility of the auditory stimuli has repeatedly been found to be greatly improved when the talker's articulating face is visible (e.g., Erber, 1969; Sumby & Pollack, 1954). The second line of research involves studies in which a clear auditory syllable (typically, with an initial stop consonant) is presented simultaneously with a visual presentation of a face articulating a different syllable (with a different, visually contrastive initial consonant). As first demonstrated by McGurk and MacDonald (1976), this often results in a striking phenomenon called the *McGurk effect* in which the resulting percept is a consonant other than the one presented auditorily (for example, an auditory /ba/ presented with a visual /ga/ is often perceived as /da/). Because the presentation of audiovisual stimuli with a cross-modal discrepancy results in a single speech percept incorporating phonetic information from both modalities, the McGurk effect is generally recognized as strong evidence for audiovisual integration in speech perception (e.g., Fowler, 1986; Green, 1998; Massaro, 1987, 1998; McGurk & MacDonald, 1976; Summerfield, 1987).

The McGurk effect is phenomenally very compelling and has been replicated many times with different stimuli under a variety of manipulations (e.g., Green & Gerdeman, 1995; Green, Kuhl, Meltzoff, & Stevens, 1991; Jordan & Bevan, 1997; MacDonald & McGurk, 1978; Manuel, Repp, Liberman, & Studdert-Kennedy, 1983; Massaro & Cohen, 1996; Rosenblum & Saldaña, 1996). However, the effect does not always occur. That is, when presented with an audiovisually discrepant stimulus (e.g., an auditory /ba/ with a visual /ga/), subjects will sometimes make an apparent "auditory" response (e.g., /ba/); that is, they identify the consonant in a manner consistent with the information in the auditory signal. To some extent, this depends on the specific audiovisual pairing, since certain audiovisual configurations give rise to a stronger McGurk effect than others. For example, a visual /ga/ tends to produce a fairly robust McGurk effect with an auditory /ba/, whereas a visual /ga/ almost never affects identification of an auditory /da/, presumably be-

This research was supported by NIH postdoctoral fellowship F32 DC00373 awarded to L.B. and NIH Grant R01 DC00130 awarded to J. L.M. Preparation of the manuscript was also supported by NIH Grant HD01994, awarded to Haskins Laboratories. We thank Matthew Paré, Sarah Whiting, Heather Berg, and Sean Toner for help in preparing the stimuli and testing subjects. Correspondence should be addressed to L. Brancazio, Department of Psychology, Southern Connecticut State University, 501 Crescent St., New Haven, CT 06515 (e-mail: brancazioL1@southernct.edu).

cause /ga/ is visually similar to /da/ and thus does not produce a sufficient audiovisual discrepancy. Generally, however, it is not uncommon to find that for an audiovisual stimulus that does generate a McGurk effect, repeated presentations of the stimulus over the course of an experiment will result in a McGurk effect on only a percentage of trials (e.g., Brancazio, 2004; Green & Norrix, 1997; MacDonald & McGurk, 1978; Massaro & Cohen, 1983). Furthermore, the overall incidence of the McGurk effect for a given stimulus typically varies considerably across individuals (Brancazio, Miller, & Paré, 1999; Carney, Clement, & Cienkowski, 1999; MacDonald, Andersen, & Bachmann, 2000).

Accordingly, a full understanding of the McGurk effect requires an explanation of why, for stimuli that typically produce the effect, it sometimes fails to occur. In this article, we will refer to such occurrences as *non-McGurk* responses. (Note that our definition of non-McGurk responses does not include responses to audiovisually incongruent stimuli that do not typically give rise to a McGurk effect, such as an auditory /da/ paired with a visual /ga/, which is almost universally identified as /da/.) One immediately apparent explanation for the incidence of non-McGurk responses is that the phonetically relevant information in the visual signal is not being used in those cases and that the resulting phonetic percept is solely determined on the basis of the auditory stimulus. However, an alternative explanation for the occurrence of non-McGurk responses is that they, like McGurk responses, arise as a consequence of the perceptual integration of the discrepant auditory and visual stimuli, but that the resulting percept is closest to the consonant specified in the auditory signal. For example, when an auditory /ba/ is presented with a visual /da/, integration results in a percept that lies somewhere between two (or more) phonetic categories (between /b/, /d/, and possibly /ð/). In this case, the percept is ultimately mapped by the perceiver onto one phonetic category, and on some trials, the stimulus may be judged to be a better match to the category for /b/ than to the categories for /d/ or /ð/, resulting in a non-McGurk percept (Brancazio, 2004; Massaro, 1987, 1998). According to this explanation, a non-McGurk percept does not reflect a failure to integrate the information in the two modalities.

The question of whether non-McGurk responses reflect a failure to use phonetically relevant information in the visual modality is especially relevant in light of the fact that the overall incidence of the McGurk effect is often used as an index of the influence of visual information in speech perception. For example, Sekiyama and colleagues (Sekiyama, 1997; Sekiyama & Tohkura, 1991, 1993), in a series of cross-cultural studies of the McGurk effect, found a lower incidence of McGurk responses by native speakers of Japanese and Chinese, compared with native speakers of English. Sekiyama and colleagues interpreted the lower incidence of McGurk responses as evidence that native Japanese and Chinese speakers (perhaps for cultural reasons) rely less on visual information when perceiving speech than do native English speakers and moreover that they "use a type of processing in which visual information is not integrated with auditory information" (Sekiyama, 1997, p. 74; but see Massaro, Cohen, Gesi, Heredia, & Tsuzaki, 1993, and Massaro, Cohen, & Smeele, 1995). However, the lower incidence of McGurk responses by these individuals might not be due to a failure to use phonetic information in the visual signal. Instead, perhaps native Japanese and Chinese speakers use visual information to the same extent as native English speakers in perceiving speech but differ in the phonological space onto which the resulting integrated percepts are mapped (Burnham, 1998; Chen & Massaro, 2004; Massaro et al., 1993). In other words, an integrated percept arising from, for example, an auditory /b/ and visual /d/ might be closer to the /d/ category in English phonological space, but closer to /b/ in Japanese and Chinese phonological spaces. (In line with this possibility, previous work has demonstrated that categorization of audiovisually discrepant stimuli can be highly influenced by the phonological inventory of the perceiver's native language; Burnham, 1998; Werker, Frost, & McGurk, 1992.)

One way to address the question of whether a failure to obtain a McGurk effect necessarily reflects a failure to use visual information in phonetic processing is to directly examine instances of non-McGurk responses. To date, however, very few studies have investigated the nature of the percepts in these cases. MacDonald et al. (2000) presented audiovisually discrepant stimuli that differed in the level of degradation of the visual stimulus and asked subjects to identify the initial consonant and to rate the auditory clarity of each stimulus. They found that for a subset of their subjects who produced mostly non-McGurk responses, the auditory clarity ratings were lower for sounds presented with intact visual stimuli than for those presented with highly degraded visual stimuli, indicating that these perceivers were influenced by the available visual information. In a similar vein, Brancazio (2004) presented subjects with audiovisually congruent and incongruent stimuli and asked them to identify the initial consonant and to rate the goodness of the percept as an exemplar of the selected consonant. Notably, ratings were lower for stimuli in the incongruent condition when the McGurk effect did not occur (e.g., auditory /b/ and visual /d/ perceived as /b/) than for audiovisually congruent stimuli, and for McGurk responses (e.g., auditory /b/ and visual /d/ perceived as /d/) to the same incongruent stimuli. In a reaction time (RT) study using the same stimuli, Brancazio also found that classification times for consonants were slower for non-McGurk responses to the incongruent stimuli than for correct responses to the congruent stimuli. The low goodness ratings and slow categorization response times for non-McGurk responses, along with the low auditory clarity ratings reported by MacDonald et al., indicate that the visual stimulus was not altogether disregarded when the McGurk effect did not occur; clearly, perceivers were affected by the discrepancy. This is consistent with the notion that audiovisual integration can result in non-

McGurk percepts. On this view, the auditory clarity and category goodness ratings for non-McGurk percepts were low, and the RTs for these percepts were slow, because in such cases the integrated percept was a poor match to the selected phonetic category. However, it is also possible that the low clarity and goodness ratings and slow RTs occurred because subjects attended to both the auditory and visual modalities without integrating the information from the two and detected the audiovisual discrepancy, which caused a disruption in processing (thus, causing slower RTs) and induced a response bias toward lower clarity and goodness ratings. Thus, neither MacDonald et al.'s nor Brancazio's results provide conclusive evidence that visual information is used in phonetic processing (that is, integrated with auditory information) in instances when the McGurk effect fails to occur.

The purpose of the present study was to explore further the nature of non-McGurk percepts in the face of audiovisual discrepancy, and specifically to determine whether perceivers are using any visual phonetically relevant information when they fail to experience a McGurk effect. We addressed this question by exploiting a different phonetic consequence of visual speech than the McGurk effect. It is important to note that although the McGurk effect involves a change in perceived place of articulation arising from place information in the visual signal, the processing of phonetic dimensions other than place of articulation can also be influenced by information in the visual signal. One notable example was reported by Green and Miller (1985), who found that the perception of voicing is influenced by visual speaking rate. They presented stimuli from a voice-onset-time (VOT) continuum ranging from /bi/ to /pi/ with visual stimuli of a person speaking a bilabial-initial syllable (either /bi/ or /pi/) quickly or slowly. They found that the /b/–/p/ voicing boundary occurred at a shorter VOT with fast videos than with slow videos, paralleling the previously established effect of auditory speaking rate (specified by syllable length) on voicing boundaries (Miller, 1981; Summerfield, 1981). Given the lack of discrepancy in place of articulation between the auditory and visual stimuli, Green and Miller's finding provides an example of a visual phonetic effect independent of place of articulation.

In the present study, we used this visual rate effect to test whether perceivers are using any phonetically relevant information in the visual signal when they fail to exhibit a McGurk effect. We created a set of stimuli designed to produce both a McGurk effect (on some, but not all, trials) and a visual rate effect, in order to test whether the visual rate effect would occur on trials when the McGurk effect fails to occur. Such a finding would provide evidence that a failure to obtain a McGurk effect does not necessarily indicate a failure to use visual information in phonetic processing.

Specifically, we presented stimuli from a /bi/–/pi/ auditory continuum with fast and slow visual /ti/ stimuli. In doing so, we borrowed a strategy first used by Green and Kuhl (1989) to study the effects of visual place of articulation on voicing boundaries—namely, pairing an auditory bilabial continuum with a visual stimulus that has a different place of articulation. Green and Kuhl presented stimuli from an auditory /ibi/–/ipi/ continuum either unimodally or with a visual /igi/ that almost always produced a McGurk effect (the stimuli were perceived as either /idi/ or /iti/ on 97% of the trials); they found that when the bilabial auditory stimuli were perceived as alveolar due to the visual stimulus, the voicing boundary occurred at a longer VOT than when there was no visual stimulus, comparable to the established effect of auditory place of articulation on voicing boundaries (Lisker & Abramson, 1970; Miller, 1977). More recently, we (Brancazio et al., 1999) modified this technique by using stimuli that produced substantial numbers of both McGurk and non-McGurk responses (specifically, visual /ti/ in conjunction with an auditory /bi/–/pi/ continuum), in order to compare voicing boundaries for McGurk and non-McGurk responses to the same stimuli. That is, we obtained both /d/ (McGurk) and /b/ (non-McGurk) responses at short VOTs, both /t/ (McGurk) and /p/ (non-McGurk) responses at long VOTs, and all four responses at intermediate VOTs. The latter approach allows the computation of voicing boundaries both when the McGurk effect occurs (i.e., a /d/–/t/ boundary) and when it does not occur (i.e., a /b/–/p/ boundary) for a single set of audiovisually discrepant stimuli. In line with Green and Kuhl's findings, we found that the voicing boundary for non-McGurk (/b/–/p/) responses occurred at a shorter VOT than the voicing boundary for McGurk (/d/–/t/) responses.

The present study extended the methodology used in Brancazio et al. (1999) to include a visual rate manipulation; that is, we used both fast and slow tokens of visual /ti/. We anticipated that we would be able to compute voicing boundaries for McGurk responses (/d/ and /t/) and for non-McGurk responses (/b/ and /p/) separately with the fast and slow visual stimuli. This would allow us to determine whether the rate effect occurs both when the McGurk effect occurs and when it does not. A finding of a /b/–/p/ boundary at a shorter VOT with the fast visual stimulus than with the slow visual stimulus would provide evidence for the use of visual phonetically relevant information in instances when the McGurk effect fails to occur.

We report the results of two experiments. Experiment 1 was a preliminary study intended to replicate Green and Miller's (1985) finding of a visual rate effect when the auditory and visual stimuli have the same place of articulation. Experiment 2 was the critical experiment, in which we used the same stimuli to test for a visual rate effect when the auditory and visual stimuli differ in place of articulation, allowing us to conduct separate analyses for McGurk and non-McGurk responses.

EXPERIMENT 1

The purpose of Experiment 1 was to establish the visual speaking rate effect reported by Green and Miller

(1985) with our stimuli. Because our main question (to be tested in Experiment 2) involved presentation of stimuli from an auditory /bi/–/pi/ continuum with fast and slow visual /ti/ stimuli, we needed to establish both that the stimuli in our auditory /bi/–/pi/ continuum were susceptible to a visual rate effect and that our fast and slow visual /ti/ stimuli would generate a rate effect. Accordingly, we presented stimuli from an auditory /bi/–/pi/ continuum with fast and slow visual /pi/ stimuli in one condition and stimuli from an auditory /di/–/ti/ continuum with fast and slow visual /ti/ stimuli in a second condition. (In contrast, Green & Miller only tested for a visual rate effect using bilabial stimuli.) We predicted that the /b/–/p/ voicing boundary would occur at a shorter VOT value when the fast visual /pi/ was presented than when the slow visual /pi/ was presented and likewise that the /d/–/t/ voicing boundary would occur at a shorter VOT value with presentation of the fast visual /ti/ than with presentation of the slow visual /ti/.

Method

Subjects. Eighteen members of the Northeastern University community participated in the experiment. Nine participated in the bilabial condition, and 9 participated in the alveolar condition. All were native speakers of American English between the ages of 18 and 45 who reported no speech or hearing disorders and who had normal or corrected-to-normal vision. All subjects were paid for their participation.

Stimuli. There were two sets of stimuli: a bilabial set and an alveolar set. The bilabial stimuli consisted of an auditory series of syllables varying in VOT and ranging from /bi/ to /pi/, which were presented concurrently with digitally modified fast and slow versions of a visual token of a mouth producing /pi/. Likewise, the alveolar stimuli consisted of an auditory series of stimuli varying in VOT and ranging from /di/ to /ti/, presented with fast and slow versions of a visual token of a mouth producing /ti/. The auditory and visual stimuli were saved as independent files, and cross-modal alignment was achieved online (see the Procedure section).

Auditory stimuli.The auditory stimuli were a subset of the stimuli created for an earlier study by Brancazio, Miller, and Paré (2003); a full description of the stimulus creation procedure is provided in that article. Briefly, the auditory continua were created by performing an LPC-based synthesis on an utterance of /bi/ (with a duration of 522 msec) and an utterance of /di/ (with a duration of 536 msec), spoken by a female native speaker of American English. For each original utterance (/bi/ and /di/), a series of stimuli that differed only in VOT was created by incrementally converting voiced pitch periods, starting with the first pitch period following the initial burst and aspiration, into voiceless segments of equivalent length (the first stimulus was created without converting any pitch periods; the second was created by converting the first pitch period into a voiceless segment; the third was created by converting the first two pitch periods into voiceless segments, etc.). The step size in VOT was approximately 5 msec for both series. By this technique, syllable duration was held constant across each continuum. Each stimulus was subsequently truncated to 300 msec (with a descending cosine ramp over the final 30 msec in order to simulate a realistic amplitude contour). Additionally, 833 msec of silence was added to each syllable prior to the consonant burst, so that the release burst in each syllable would be temporally aligned with that in each of the visual stimuli (see the Procedure section for audiovisual alignment technique).

A subset of 12 of the stimuli in each continuum was selected for use in the present experiment. The VOTs of the selected stimuli in the /bi/–/pi/ series were (in milliseconds): 10, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, and 74. The VOTs of the selected stimuli in the /di/–/ti/ series were (in milliseconds): 21, 31, 36, 41, 46, 51, 55, 60, 65, 70, 75, and 84. (The /bi/–/pi/ series started at a shorter VOT than the /di/–/ti/ series because the original utterance of /bi/ had a shorter VOT than the original utterance of /di/.) Thus, each experimental series involved steps of one pitch period (typically 5 msec) increments with the exception of the endpoints, which differed from their respective adjacent stimuli by two pitch periods (approximately 10 msec in each case).

Visual stimuli. Our approach for creating a visual rate manipulation was different from that of Green and Miller (1985). They recorded a speaker who produced fast and slow utterances of /bi/ and /pi/, and they used several different fast and slow utterances as visual stimuli. In contrast, we digitally edited one visual token each of /pi/ and /ti/ (originally spoken at a moderate rate) by deleting or duplicating video frames to create the impression of fast and slow speaking rates, respectively. This technique provided us with a visual speaking rate manipulation in which all aspects of the visual stimuli other than speaking rate were held constant.

The visual /pi/ and /ti/ stimuli were modified versions of stimuli created for the Brancazio et al. (2003) study; again, a full description of the creation of the original stimuli is available in that article. Briefly, the same speaker who was recorded for the auditory tokens was videotaped producing the utterances /əpi/ and /əti/, such that the recorded images included the face from just below the nose to a point just below the jawline at maximal vowel opening. The stimuli were digitized on a Macintosh G3 computer, using Adobe Premiere software, at 30 frames per second at a resolution of 320 × 240 pixels. The stimuli were then edited so that they were equal in duration, the onset occurred during the steady-state portion of the schwa (thus making the utterances appear as /pi/ and /ti/, preceded by a neutral mouth configuration), and the point of consonant release occurred in the same frame for the two stimuli. Additionally, three repetitions of the initial frame were included at the onset to prevent an abrupt stimulus onset. (The stimuli created for the Brancazio et al., 2003, study had five repetitions of the initial frame, but two of these were removed for the present study because it better suited the visual rate manipulation, described below.)

These /pi/ and /ti/ visual tokens were modified to create fast and slow versions of each. To create fast versions of the visual /pi/ and /ti/ tokens, we deleted every other frame (33 msec each) in each video clip, effectively doubling the speed. To create slow versions of the tokens, we initially duplicated every frame in each clip, effectively halving the speed. However, in preliminary tests, we determined that the consonant releases in these slow stimuli appeared unnatural. Thus, we further modified the slow stimuli by triplicating two video frames occurring immediately before the consonantal closure and two frames during the closure, and keeping all of the other frames duplicated.[1] (The initial three static frames were not modified in either the fast or slow manipulation.) As a result, the fast videos consisted of 19 video frames, and the slow videos consisted of 53 video frames (with the total including the initial three static frames in both cases).

To equate the lengths of the movie files and to make the visible consonant release occur in the same frame for all of the videos, we padded the onsets and offsets of the videos with frames that consisted of a black screen. The fast videos were padded with 18 black frames prior to the movie onset and 21 black frames after the movie offset; the slow videos were padded with 1 black frame at the beginning and 4 black frames at the end. As a consequence, each video was 58 frames (1,933 msec) long, and the point of visible consonant release occurred in the 26th frame (833 msec from stimulus onset). Finally, to make the transitions between the black screen to the face video less abrupt, we used a fade transition in Adobe Premiere over the first two frames and the last four frames of the face in each video.

Each visual stimulus was saved without an auditory channel.

Procedure. The subjects were randomly assigned to the bilabial condition, in which they were presented with the auditory /bi/–/pi/

stimuli paired with the fast and slow visual /pi/ stimuli, or to the alveolar condition, in which they were presented with the auditory /di/–/ti/ stimuli paired with the fast and slow visual /ti/ stimuli.

The stimuli were presented on a 17-in. monitor connected to a Macintosh G3 computer located in a sound-attenuated booth. The video image filled approximately three quarters of the screen and was surrounded by a black screen. The auditory stimuli were presented binaurally through Sony MDR-V6 headphones at a comfortable listening level. The subjects were seated in front of the computer monitor at a normal viewing distance (approximately 18 in.). The experimenter observed the experiment through a window in the booth to confirm that the subjects watched the visual presentations.

The stimulus presentation was controlled by PsyScope (Cohen, MacWhinney, Flatt, & Provost, 1993). Each trial in the experiment began with a warning beep and a printed message on the computer monitor instructing the subject to watch the monitor. Following the warning, an auditory stimulus and a visual stimulus were presented simultaneously. Because there was silence at the beginning of the auditory stimulus file, the face was visible prior to the acoustic onset of the syllable; the visual stimulus also remained visible after the acoustic offset (but only briefly so for the fast visual stimuli). The acoustic onset of the syllable always coincided with the visual point of consonant release. After the offset of the visual stimulus, two boxes on the monitor labeled "B" and "P" or "D" and "T," depending on the condition, appeared on the screen and disappeared after the subject responded by clicking in one of the boxes. The next trial began after the subject moved the cursor to the bottom of the screen. The screen-based responding and trial initiation procedure was intended to maintain the subjects' visual attention on the computer monitor throughout the experiment. The subjects were told that they would simultaneously hear a syllable and see a mouth produce a syllable, but that the heard and seen syllables might not match. They were instructed to both listen to the syllables and watch the mouth on the monitor, and to identify the initial consonant that they heard in each syllable, regardless of what they saw.

Each condition consisted of 13 randomized blocks of the stimuli. Each block contained the 12 auditory stimuli presented with the fast and slow visual stimuli, randomized together for a total of 24 stimuli. The first block was not included in the data analysis; thus, for the purposes of our analyses, each auditory stimulus was presented 12 times with the fast visual stimulus and 12 times with the slow visual stimulus. The subjects were given a break halfway through the experiment, which lasted approximately 30 min.

Data analysis. The dependent measure for the analyses was the location of the category boundary (in milliseconds) between voiced and voiceless stops along the VOT continuum. This was calculated in the following manner for each subject. First, the percentage of responses that involved the voiceless choice (/p/ in the bilabial condition and /t/ in the alveolar condition) was determined at each step along each auditory continuum, separately for each speaking rate. Second, the voiced–voiceless boundaries were determined by fitting a normal ogive to each of the percent-voiceless functions and calculating the mean of the ogive (that is, the 50% point on the function), corresponding to the VOT value (in milliseconds) at which voiced and voiceless responses were equally probable.[2]

Subject criteria. Each subject's data were included in the analysis if he or she labeled at least one of the two stimuli with the shortest VOTs as voiced (/b/ or /d/, depending on the continuum) and at least one of the two stimuli with the longest VOTs as voiceless (/p/ or /t/, depending on the continuum), on at least 90% of the trials in both the fast and slow conditions. All of the subjects met these criteria.

Results and Discussion

The percentage of voiceless responses as a function of VOT, place of articulation, and speaking rate are pre-

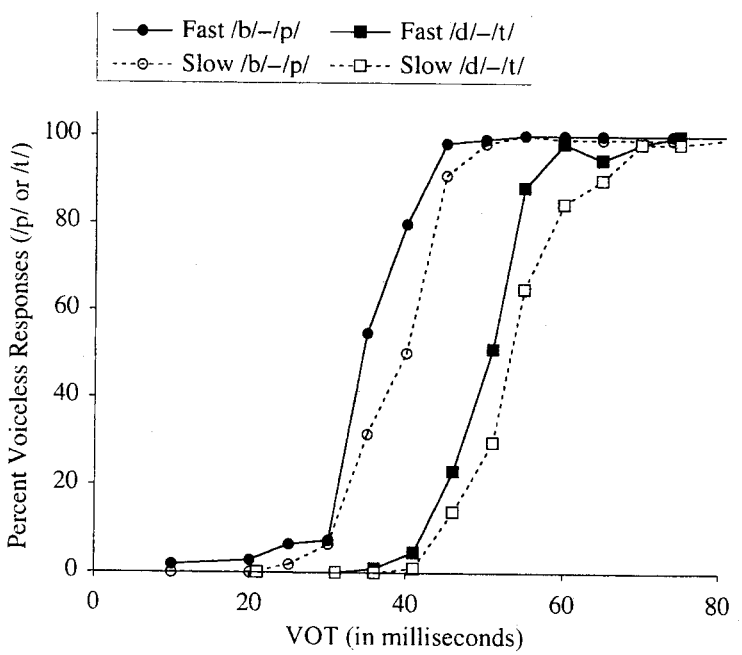

Figure 1. Group mean percentage of voiceless responses as a function of voice-onset time (VOT) in Experiment 1. The separate functions display the percentage of /p/ responses for the auditory /bi/–/pi/ continuum presented with the fast visual /pi/ and the slow visual /pi/, and the percentage of /t/ responses for the auditory /di/–/ti/ continuum presented with the fast visual /ti/ and the slow visual /ti/.

sented in Figure 1. As seen in the figure, the identification functions for the alveolar series are shifted toward longer VOT values relative to the functions for the bilabial series, as is consistent with previous findings (e.g., Lisker & Abramson, 1970; Miller, 1977). More important, the figure also demonstrates that for both the bilabial and alveolar series, the identification function in the fast condition is shifted toward shorter VOT values, relative to the slow condition. These shifts are reflected in the voicing boundaries: The /b/–/p/ boundaries in the fast and slow conditions were 35.62 msec (SD = 4.05) and 38.22 msec (SD = 4.47), respectively; the /d/–/t/ boundaries in the fast and slow conditions were 49.94 msec (SD = 3.54) and 53.58 msec (SD = 4.00), respectively. Thus, there was a 2.60-msec rate effect for the bilabial stimuli, and a 3.64-msec rate effect for the alveolar stimuli.

The voicing boundaries were analyzed in a 2 × 2 analysis of variance (ANOVA) with visual rate (fast vs. slow) as a within-subjects factor and place of articulation (bilabial vs. alveolar) as a between-subjects factor. The ANOVA revealed significant main effects of both visual rate [$F(1,16) = 34.14, p < .0001$] and place of articulation [$F(1,16) = 66.40, p < .0001$], with no interaction between the factors ($F < 1$).

Thus, Experiment 1 successfully established that our auditory /bi/–/pi/ series is susceptible to a visual rate effect, and that our fast and slow visual /ti/ stimuli can generate a visual rate effect, when the auditory and visual stimuli have the same place of articulation.

EXPERIMENT 2

As stated earlier, the purpose of Experiment 2 was to determine whether visual speech information is used in phonetic processing when the McGurk effect fails to occur, by testing for a visual rate effect in non-McGurk responses using stimuli that sometimes (but not always) generate a McGurk effect. To do so, we added an audiovisual discrepancy to the visual rate effect paradigm used in Experiment 1. Specifically, using the stimuli in Experiment 1, we presented the auditory /bi/–/pi/ continuum with the fast and slow tokens of visual /ti/. We expected subjects to provide a mixture of McGurk (/d/ and /t/) and non-McGurk (/b/ and /p/) responses. This allowed us to determine whether there would be a rate shift not only when the McGurk effect occurred (i.e., whether there would be a shift in the /d/–/t/ boundary with the fast and slow visual /ti/ stimuli), but also when it did not occur (i.e., whether there would be a shift in the /b/–/p/ boundary with the fast and slow visual /ti/ stimuli). Our critical test involved the latter of these questions: A finding of a visual rate effect for non-McGurk responses (/b/ and /p/) would provide evidence that perceivers use visual phonetically relevant information in speech perception in the absence of a McGurk effect.

There were two notable methodological differences between Experiments 1 and 2, aside from the use of incongruent visual stimuli. As we will explain in the Method section, we partitioned responses according to whether or not a McGurk effect occurred and conducted separate analyses of McGurk and non-McGurk responses. As a result, it was necessary to include many more repetitions of each stimulus than in Experiment 1. In addition, many subjects provided data that permitted an analysis only of non-McGurk responses or only of McGurk responses, necessitating a much larger sample size than in Experiment 1.

Method

Subjects. Forty-eight members of the Northeastern University community participated in the experiment, none of whom had participated in Experiment 1. All were native speakers of American English between the ages of 18 and 45 who reported no speech or hearing disorders, and who had normal or corrected-to-normal vision. All the subjects were paid for their participation.

Stimuli. The stimuli used for this experiment were the auditory /bi/–/pi/ continuum and the fast and slow visual /ti/ stimuli used in Experiment 1.

Procedure. The experiment consisted only of an audiovisually incongruent condition, in which each of the auditory stimuli from the /bi/–/pi/ continuum was presented with the fast and slow visual /ti/ stimuli. The experiment was conducted over two sessions held on separate days. Each session lasted approximately 45 min and consisted of 19 randomized blocks of the stimuli; as in Experiment 1, each block consisted of the 12 auditory stimuli presented with the fast and slow visual stimuli, randomized together for a total of 24 stimuli. The first block was not included in the data analysis, so that for the purposes of our analyses, each auditory stimulus was presented 36 times with each video, for a total of 864 trials. In other respects, the experiment was conducted in the same manner as the previous one, except that on each trial, the subjects were given the choices "B," "P," "D," and "T" for their response.

Data analysis. As in Experiment 1, the dependent variable was the location of the voicing boundary, measured in VOT. Boundaries were calculated separately for the /d/ and /t/ (McGurk) responses and for the /b/ and /p/ (non-McGurk) responses, independently for each visual speaking rate condition, using the following steps. First, all of the responses for each subject were partitioned according to whether they were bilabial (/b/ or /p/) or alveolar (/d/ or /t/). Second, the percentage of voiceless responses (/p/ or /t/) was calculated for each continuum step, separately for the fast and slow conditions. Finally, the voicing boundary in each case was calculated by fitting a normal ogive to the percent-voiceless function. As a result, for each subject, we calculated (when possible; see below) /b/–/p/ and /d/–/t/ boundaries in the fast and slow conditions.

Subject criteria. There was considerable variation in the magnitude of the McGurk effect across subjects; some subjects provided few or no non-McGurk (/b/ or /p/) responses at certain steps along the continuum, whereas others provided few or no McGurk (/d/ or /t/) responses at certain steps. To ensure that the voicing boundaries were meaningful, we only included a subject's /b/–/p/ and /d/–/t/ voicing boundaries in the analyses if certain criteria were met. These criteria were applied separately for each subject's /b/–/p/ and /d/–/t/ data, meaning that a given subject might provide a usable /b/–/p/ boundary but not a usable /d/–/t/ boundary, or vice versa.

The following criteria had to be met in both the fast and slow conditions for /b/–/p/ boundaries to be calculated. First, the subject had to respond with either /b/ or /p/ on at least 20% of all trials, averaged across the continuum. Second, the endpoints had to be consistently identified, such that the subject labeled at least one of the first two steps (i.e., with the shortest VOTs) as /b/ on at least 90% of the trials and labeled at least one of the final two steps (i.e., with

the longest VOTs) as /p/ on at least 90% of the trials. Third, each continuum step in the ambiguous region had to have at least five responses (either /b/ or /p/); the ambiguous region was determined individually for each subject as a region of continuum steps bounded by a step that received 100% /b/ responses (or the continuum endpoint if no step had 100% /b/ responses) and a step that received 100% /p/ responses (or the endpoint if no step had 100% /p/ responses). We applied equivalent criteria for /d/ and /t/ responses to determine whether /d/–/t/ boundaries would be calculated for each subject.

Results and Discussion

Overall, subjects made a McGurk response (either /d/ or /t/) on 48.3% of the trials, averaged across all of the stimuli. However, there was considerable variability in the magnitude of the McGurk effect across subjects, ranging from 0% to 100%. Of the 48 subjects, 32 provided enough McGurk responses (/d/ or /t/) to allow computation of /d/–/t/ boundaries in the fast and slow conditions (based on the criteria described earlier), and 27 provided enough non-McGurk responses (/b/ or /p/) to allow computation of /b/–/p/ boundaries. Fifteen of these subjects provided enough McGurk and non-McGurk responses to allow computation of both /b/–/p/ and /d/–/t/ boundaries.

First, we examined whether there was a visual rate effect on trials in which the McGurk effect occurred—that is, whether there was a difference in the /d/–/t/ boundaries in the fast and slow conditions. For the 32 subjects who provided calculable /d/–/t/ boundaries, the voicing boundary occurred at a shorter VOT in the fast condition than in the slow condition: The mean /d/–/t/ boundary in the fast condition was 39.45 msec VOT (SD = 3.97), whereas the mean /d/–/t/ boundary in the slow condition was 41.39 msec VOT (SD = 4.23). A paired t test revealed that this 1.94-msec difference was significant [$t(31) = 7.27, p < .0001$].

Next, we addressed the critical question of whether a visual rate effect would also emerge on trials in which the McGurk effect did not occur—that is, whether there was a difference in the /b/–/p/ boundaries in the fast and slow conditions. As with the /d/–/t/ responses, there was a small shift in the voicing boundary to a shorter VOT in the fast condition. For the 27 subjects who provided calculable /b/–/p/ boundaries, the mean /b/–/p/ boundary was 32.28 msec VOT (SD = 3.58) in the fast condition and 33.24 msec VOT (SD = 3.92) in the slow condition. A paired t test revealed this .96-msec difference to be statistically significant [$t(26) = 3.69, p = .001$].

We noted that the magnitude of the rate effect for non-McGurk responses (.96 msec) was somewhat smaller than that of the effect for McGurk responses (1.94 msec). Accordingly, we conducted an additional analysis to directly compare the rate effects for McGurk and non-McGurk responses, using the data from the 15 subjects who provided both /b/–/p/ and /d/–/t/ boundaries (note that these subjects constitute a subset of the subjects who provided data for the separate McGurk and non-McGurk analyses reported above). Figure 2 presents, for these 15 subjects, the percentage of voiceless responses as a func-

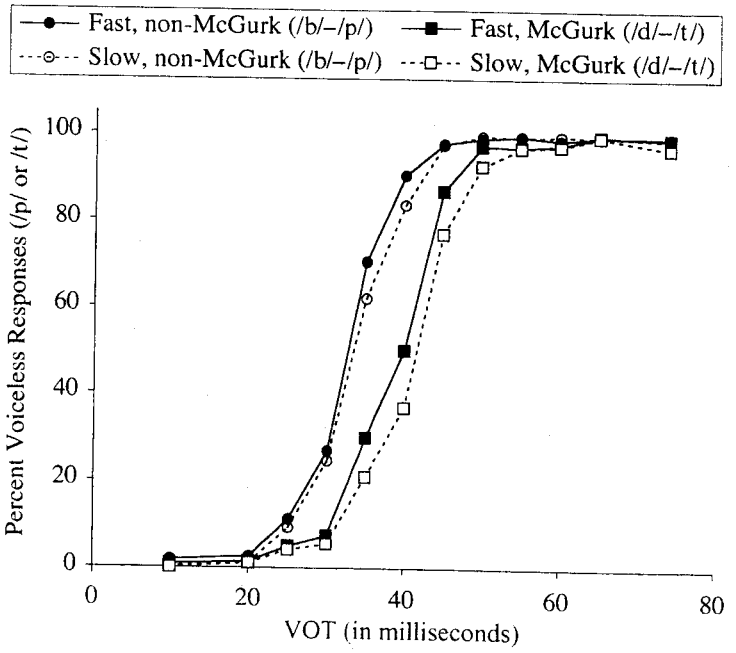

Figure 2. Group mean percentage of voiceless responses as a function of voice-onset time (VOT) in Experiment 2. The separate functions display the percentage of non-McGurk (/b/ or /p/) responses that were /p/, and the percentage of McGurk (/d/ or /t/) responses that were /t/, when the auditory /bi/–/pi/ continuum was presented with the fast and slow visual /ti/.

tion of VOT for /d/ and /t/ (McGurk) and /b/ and /p/ (non-McGurk) responses in the fast and slow conditions. The figure displays the rate effects for both McGurk and non-McGurk responses, as the /d/–/t/ and /b/–/p/ identification functions in the fast condition are shifted toward shorter VOT values, relative to their counterparts in the slow condition. In addition, the figure displays a shift between the /d/–/t/ and /b/–/p/ functions, with the latter shifted toward short VOT values. The figure also illustrates the somewhat smaller rate shift for the non-McGurk responses than for the McGurk responses. The mean /d/–/t/ boundaries in the fast and slow conditions were 38.94 msec VOT (SD = 3.79) and 40.90 msec VOT (SD = 4.01), respectively; the mean /b/–/p/ boundaries in the fast and slow conditions were 32.61 msec VOT (SD = 2.83) and 33.75 msec VOT (SD = 3.44), respectively.

A 2 × 2 ANOVA with visual rate (fast vs. slow) and perceived place of articulation (alveolar vs. bilabial) as within-subjects factors revealed significant main effects of visual rate [$F(1,14) = 28.92, p < .0001$] and perceived place of articulation [$F(1,14) = 71.26, p < .0001$]. (Note that the place of articulation effect replicates Brancazio et al.'s [1999] finding, using similar stimuli, that the /b/–/p/ boundary for non-McGurk responses falls at a shorter VOT than does the /d/–/t/ boundary for McGurk responses with a single set of stimuli.) The interaction between visual rate and perceived place of articulation was marginal [$F(1,14) = 4.00, p = .07$], suggesting a tendency toward a stronger rate effect for McGurk responses (1.96 msec) than for non-McGurk responses (1.14 msec). However, separate t tests for the McGurk and non-McGurk voicing boundaries revealed that the rate effect was significant for both sets of responses [McGurk: $t(14) = 5.36, p < .0001$; non-McGurk: $t(14) = 3.34, p < .005$], which is consistent with the analyses involving the larger sets of subjects. All 15 subjects showed an effect in the expected direction (that is, the fast boundary occurred at a shorter VOT than did the slow boundary) for McGurk responses, and 13 of the 15 subjects showing an effect in the expected direction for non-McGurk responses. (Interestingly, only 11 of the 15 subjects exhibited a numerically larger rate effect for McGurk than non-McGurk responses, which is consistent with the finding of only a marginally significant interaction of rate and perceived place of articulation.)[3]

One question regarding this analysis is whether the subset of 15 subjects is representative of the larger sample (n = 27) of subjects who provided /b/–/p/ boundaries and the larger sample (n = 32) of subjects who provided /d/–/t/ boundaries. To address this, we conducted a t test to compare the rate effect for non-McGurk responses (measured as the difference, in VOT, between the /b/–/p/ boundaries in the fast and slow conditions) for the 15 subjects who provided both /b/–/p/ and /d/–/t/ boundaries and for the remaining 12 subjects (from the set of 27) who provided only /b/–/p/ boundaries, and found that the difference was not significant [$t(25) = .76, p > .10$]. We also conducted a t test to compare the rate effect for McGurk responses (the difference between the /d/–/t/ boundaries in the fast and slow conditions) for the 15 subjects who provided /b/–/p/ and /d/–/t/ boundaries and for the remaining 17 (from the set of 32) who provided only /d/–/t/ boundaries; here, too, the difference was not significant [$t(30) = .06, p > .10$]. Thus, the subset of subjects did not differ from the larger sample on the magnitude of the visual rate effect.

Overall, our results indicate a visual rate effect for non-McGurk responses. However, this effect was somewhat smaller in magnitude than the visual rate effect for McGurk responses. A possible reason for this is that a subset of subjects failed to use the visual rate information when they did not experience a McGurk effect. As we noted earlier, there was considerable variability in the overall incidence of the McGurk effect across subjects; even among subjects who provided non-McGurk results, the incidence of the McGurk effect ranged from 0% to 63.5%. The subjects who rarely experienced the McGurk effect might have used visual information less (and thus did not show a rate effect for non-McGurk responses) than those who exhibited the McGurk effect somewhat more often.

To address this possibility, we divided the 27 subjects who provided /b/–/p/ (non-McGurk) results into three equal-sized groups (n = 9), on the basis of each subject's overall percentage of McGurk responses, averaged across all of the stimuli. The percentage of McGurk responses ranged from 0% to 16% for the "low" group, from 25% to 50% for the "medium" group, and from 51% to 63% for the "high" group. We conducted a 2 × 3 ANOVA with visual rate (fast vs. slow) as a within-subjects factor and McGurk effect magnitude (low vs. medium vs. high) as a between-subjects factor. The effect of visual rate was significant [$F(1,24) = 12.81, p < .005$], as expected, and there was no main effect of McGurk effect magnitude ($F < 1$; indicating that the three subgroups did not differ in their overall boundary locations, independent of rate). More important, the interaction between visual rate and McGurk effect magnitude was not significant ($F < 1$). Thus, there is no evidence to indicate that subjects who exhibited an especially weak McGurk effect were any less susceptible to the visual rate effect for non-McGurk percepts than were subjects who experienced a stronger McGurk effect.

In summary, the important finding of Experiment 2 was that, for stimuli that often produced a McGurk effect, there was a significant visual rate effect not only when the McGurk effect occurred, but also when the McGurk effect did not occur.[4]

GENERAL DISCUSSION

At the beginning of this article, we outlined two reasons why the McGurk effect might sometimes fail to occur for stimuli that typically produce the effect. One

possible reason for a failure to obtain a McGurk effect is that the perceiver does not use phonetically relevant information in the visual signal, and the resulting percept is based only on information in the auditory signal. The other possible reason is that the perceiver does, in fact, use phonetically relevant information in the visual signal, but the resulting integrated percept falls closest to the phonetic category specified by the auditory stimulus alone.

Our finding in Experiment 2 of a significant visual rate effect for non-McGurk responses demonstrates that perceivers do use phonetically relevant information in the visual signal in instances when they fail to experience a McGurk effect. This finding is consistent with the claim that a failure to obtain a McGurk effect does not indicate a failure to use phonetically relevant information in the visual signal in speech perception, but, to the contrary, audiovisual integration can result in non-McGurk, ostensibly "auditory" percepts (Brancazio, 2004).

As noted earlier, audiovisual integration of an audiovisually discrepant stimulus might lead to non-McGurk responses if integration gives rise to a percept that, because of the conflicting information in the two modalities, does not perfectly match the phonetic categories specified either just auditorily or just visually, instead falling somewhere between these categories along a perceptual metric (which could take a variety of forms, including an articulatory one; Schwartz, Robert-Ribes, & Escudier, 1998; Summerfield, 1987). If the perceiver judges this intermediate percept to be "closer" to the category that is specified by the auditory stimulus alone, the result is a non-McGurk response.

This account is consistent with the results of the present study. To see why, consider the information provided in each modality in the stimuli of Experiment 2. First, there is conflicting information for place of articulation in the auditory (bilabial) and visual (alveolar) signals. Second, the auditory signal provides information about voicing (which varies across the stimuli with changes in VOT), but there is no audiovisual conflict with regard to voicing because the visual signal does not convey voicing information. Finally, both the auditory and visual modalities provide information about speaking rate; note, however, that rate varies between the fast and slow visual stimuli but is held constant across the auditory continuum. Given these sources of information, perception of the stimuli presented in Experiment 2 might proceed as follows. On a given trial, audiovisual integration of the bilabial auditory and alveolar visual stimuli results in a perceived place of articulation somewhat intermediate (along some perceptual metric) between bilabial and alveolar, which is then mapped onto the place category (either bilabial or alveolar) judged to be closest. Additionally, the VOT information from the auditory signal is evaluated in order to determine voicing, and it is well established that evaluation of VOT depends on the contextual variables of speaking rate (e.g., Miller, 1981; Summerfield, 1981) and place of articulation (e.g., Lisker & Abramson, 1970), which is consistent with rate-dependent and place-dependent differences in VOT in speech production (Lisker & Abramson, 1964; Volaitis & Miller, 1992). With the present stimuli, VOT is therefore evaluated with respect to both speaking rate (auditory and visual, although only the latter varies across stimuli) and the place of articulation category (bilabial or alveolar) onto which the intermediate percept is mapped. Ultimately, this process results in categorization of the stimulus as either voiced (i.e., /b/ or /d/) or voiceless (i.e., /p/ or /t/). Such a process would give rise to the observed visual rate effect for both McGurk and non-McGurk responses, with more voiceless responses (/t/ for McGurk responses and /p/ for non-McGurk responses) with the fast than the slow visual stimulus. It would also give rise to the finding that the voicing boundary for non-McGurk /b/ and /p/ responses occurred at a shorter VOT than for McGurk /d/ and /t/ responses.

It should be noted that Massaro's (1987, 1998) fuzzy logical model of perception (FLMP) also incorporates the assumption that audiovisual integration may give rise to non-McGurk percepts and is therefore also compatible with the present findings. According to the FLMP, auditory and visual features are independently evaluated with reference to the incoming auditory and visual signals, and stored prototypes of different syllables (with auditory and visual feature values) are then compared with the resulting feature values; the prototype with the highest degree of match is selected to arrive at a phonetic decision. In the case of an audiovisually incongruent stimulus, several prototypes will partially match the feature values; for example, for an auditory /ba/ presented with a visual /da/, the prototypes for a number of syllables, including /ba/, /da/, and /ða/, will all have partial matches to the input. Because the model is probabilistic, /ba/, /da/, /ða/, and other prototypes may each sometimes be selected as the best-matching prototype. Given this, /ba/ (the response consistent with the auditory stimulus) will sometimes occur even though audiovisual integration has occurred. To the extent that the FLMP can incorporate visual speaking rate in the feature values for voiced and voiceless consonants, it should also be able to account for the finding of a visual rate effect for non-McGurk percepts.

One question regarding the use of visual speech information in phonetic processing is whether it is obligatory—that is, whether perceivers, when presented with audiovisual speech, always make use of the available phonetic information in both modalities, even when the McGurk effect does not occur. Although the present results do provide evidence for the use of phonetically relevant visual rate information for non-McGurk responses, they cannot resolve the issue of whether such use is obligatory. In particular, the finding that the visual rate effect was somewhat smaller for non-McGurk responses than for McGurk responses leaves open the possibility that perceivers fail to use rate information in the

visual signal in some instances that result in non-McGurk responses. Unfortunately, our index of whether there was a visual influence—a shift in the voicing boundary due to visual speaking rate—cannot be evaluated on a trial-by-trial basis, so it is impossible to partition (as we have done with McGurk and non-McGurk responses) instances in which visual rate affects phonetic categorization and when it does not. Thus, it remains possible that some non-McGurk responses arise from a failure to use any phonetically relevant information in the visual signal.

However, although we cannot make strong claims regarding whether phonetically relevant visual information is always used in speech perception, our results provide evidence for a more general claim—namely, that in many (if not all) instances when the McGurk effect does not occur, perceivers have used visual information in phonetic processing. An important implication of this conclusion is that the overall incidence of the McGurk effect, if taken as an index of the use of visual information in phonetic processing, will generally underestimate the extent of visual contributions to speech perception.

Consequently, our findings have implications for how individual differences in the incidence of the McGurk effect should be interpreted. Recall that in Experiment 2, we found no relationship between the magnitude of the McGurk effect and that of the visual rate effect for non-McGurk responses for a given subject. This outcome suggests that individuals who rarely experience a McGurk effect might not actually make less use of phonetically relevant information in the visual signal than do individuals who are more susceptible to the effect. Instead, it is possible that individual differences in the magnitude of the McGurk effect may reflect differences in how the percepts are mapped onto phonetic categories, more than differences in the degree of audiovisual integration.

Finally, our results also have implications for the investigation of cross-cultural differences in the McGurk effect. Because our study only involved native English speakers, our results do not directly address the question of whether native Japanese and Chinese speakers rely less on visual information in speech perception than do native English speakers (as suggested by Sekiyama, 1997, and by Sekiyama & Tohkura, 1991, 1993). However, our results suggest that the findings that native Japanese (Sekiyama & Tohkura, 1991, 1993) and Chinese (Sekiyama, 1997) speakers rarely experience the McGurk effect should be interpreted with caution, given that they do not necessarily indicate that these individuals fail to integrate speech information across the auditory and visual modalities. We believe that our study offers a promising approach for evaluating potential cross-cultural differences in the use of visual speech information: Speakers of different languages could be tested with the paradigm we have presented here, to determine whether they differ not only in the overall magnitude of the McGurk effect, but also in the magnitude of the visual rate effect for non-McGurk responses.

REFERENCES

Brancazio, L. (2004). Lexical influences in audiovisual speech perception. *Journal of Experimental Psychology: Human Perception & Performance*, **30**, 445-463.

Brancazio, L., Miller, J. L., & Paré, M. A. (1999). Perceptual effects of place of articulation on voicing for audiovisually discrepant stimuli. *Journal of the Acoustical Society of America*, **106**, 2270.

Brancazio, L., Miller, J. L., & Paré, M. A. (2003). Visual influences on the internal structure of phonetic categories. *Perception & Psychophysics*, **65**, 591-601.

Burnham, D. (1998). Language specificity in the development of auditory–visual speech perception. In R. Campbell, B. Dodd, and D. Burnham (Eds.), *Hearing by eye II: Advances in the psychology of speechreading and auditory–visual speech* (pp. 27-60). Hove, U.K.: Psychology Press.

Carney, A. E., Clement, B. R., & Cienkowski, K. M. (1999). Talker variability effects in auditory–visual speech perception. *Journal of the Acoustical Society of America*, **106**, 2270.

Chen, T. H., & Massaro, D. W. (2004). Mandarin speech perception by ear and eye follows a universal principle. *Perception & Psychophysics*, **66**, 820-836.

Cohen, J., MacWhinney, B., Flatt, M., & Provost, J. (1993). PsyScope: An interactive graphic system for designing and controlling experiments in the psychology laboratory using Macintosh computers. *Behavior Research Methods, Instruments, & Computers*, **25**, 257-271.

Erber, N. P. (1969). Interaction of audition and vision in the recognition of oral speech stimuli. *Journal of Speech & Hearing Research*, **12**, 423-425.

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

Green, K. P. (1998). The use of auditory and visual information during phonetic processing: Implications for theories of speech perception. In R. Campbell, B. Dodd, & D. Burnham (Eds.), *Hearing by eye II: Advances in the psychology of speechreading and auditory–visual speech* (pp. 3-25). Hove, U.K.: Psychology Press.

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

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

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

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

Green, K. P., & Norrix, L. W. (1997). Acoustic cues to place of articulation and the McGurk effect: The role of release bursts, aspiration, and formant transitions. *Journal of Speech & Hearing Research*, **40**, 646-665.

Jordan, T. R., & Bevan, K. (1997). Seeing and hearing rotated faces: Influences of facial orientation on visual and audiovisual speech recognition. *Journal of Experimental Psychology: Human Perception & Performance*, **23**, 388-403.

Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, **20**, 384-422.

Lisker, L., & Abramson, A. S. (1970). The voicing dimension: Some experiments in comparative phonetics. In *Proceedings of the Sixth International Congress of Phonetic Sciences* (pp. 563-567). Prague: Academia.

MacDonald, J., Andersen, S., & Bachmann, T. (2000). Hearing by eye: How much spatial degradation can be tolerated? *Perception*, **29**, 1155-1168.

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

Manuel, S. Y., Repp, B. H., Liberman, A. M., & Studdert-Kennedy, M.

(1983, November). *Exploring the "McGurk effect."* Paper presented at the 24th meeting of the Psychonomic Society, San Diego.

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

MASSARO, D. W. (1998). *Perceiving talking faces: From speech perception to a behavioral principle*. Cambridge, MA: MIT Press.

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

MASSARO, D. W., & COHEN, M. M. (1996). Perceiving speech from inverted faces. *Perception & Psychophysics*, **58**, 1047-1065.

MASSARO, D. W., COHEN, M. M., GESI, A., HEREDIA, R., & TSUZAKI, M. (1993). Bimodal speech perception: An examination across languages. *Journal of Phonetics*, **21**, 445-478.

MASSARO, D. W., COHEN, M. M., & SMEELE, P. M. T. (1995). Cross-linguistic comparisons in the integration of visual and auditory speech. *Memory & Cognition*, **23**, 113-131.

MCGURK, H., & MACDONALD, J. (1976). Hearing lips and seeing voices. *Nature*, **264**, 746-748.

MILLER, J. L. (1977). Nonindependence of feature processing in initial consonants. *Journal of Speech & Hearing Research*, **20**, 519-528.

MILLER, J. L. (1981). Effects of speaking rate on segmental distinctions. In P. D. Eimas & J. L. Miller (Eds.), *Perspectives on the study of speech* (pp. 39-74). Hillsdale, NJ: Erlbaum.

PITT, M. A. (1995). The locus of the lexical shift in phoneme identification. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **21**, 1037-1052.

PITT, M. A., & SAMUEL, A. G. (1993). An empirical and meta-analytic evaluation of the phoneme identification task. *Journal of Experimental Psychology: Human Perception & Performance*, **19**, 699-725.

ROSENBLUM, L. D., & SALDAÑA, H. M. (1996). An audiovisual test of kinematic primitives for visual speech perception. *Journal of Experimental Psychology: Human Perception & Performance*, **22**, 318-331.

SCHWARTZ, J.-L., ROBERT-RIBES, J., & ESCUDIER, P. (1998). Ten years after Summerfield: A taxonomy of models for audio-visual fusion in speech perception. In R. Campbell, B. Dodd, and D. Burnham (Eds.), *Hearing by eye II: Advances in the psychology of speechreading and auditory–visual speech* (pp. 85-108). Hove, U.K.: Psychology Press.

SEKIYAMA, K. (1997). Cultural and linguistic factors in audiovisual speech processing: The McGurk effect in Chinese subjects. *Perception & Psychophysics*, **59**, 73-80.

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.

SEKIYAMA, K., & TOHKURA, Y. (1993). Inter-language differences in the influence of visual cues in speech perception. *Journal of Phonetics*, **21**, 427-444.

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. (1981). Articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology: Human Perception & Performance*, **7**, 1074-1095.

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). Hillsdale, NJ: Erlbaum.

VOLAITIS, L. E., & MILLER, J. L. (1992). Phonetic prototypes: Influence of place of articulation and speaking rate on the internal structure of voicing categories. *Journal of the Acoustical Society of America*, **92**, 723-735.

WERKER, J. F., FROST, P. E., & MCGURK, H. (1992). La langue et les lèvres: Cross-language influences on bimodal speech perception. *Canadian Journal of Psychology*, **46**, 551-568.

NOTES

1. Thus, the fast and slow manipulations are not equivalent: The fast manipulation doubled the speed of the original utterance, whereas the slow manipulation halved the speed for part of the utterance and reduced the speed by two thirds for other parts of the utterance. However, our goal was not to examine the effects of specific editing techniques on perceived speaking rate, but only to produce stimuli that varied as much as possible in perceived speaking rate while appearing reasonably realistic.

2. To confirm the appropriateness of fitting a normal ogive to our data, we calculated r for each subject's percent-voiceless function in each condition in Experiments 1 and 2. Overall, the fit was very good: The mean value of r (averaged across all conditions of Experiments 1 and 2) was .99, with individual values of r ranging from .95 to 1.00.

3. An alternative approach for measuring contextual effects (including rate effects) to the ogive-based analysis we have employed is to determine an ambiguous region along the continuum, and to calculate the area between the functions (by computing the percentage of responses in a given category) in this region (Pitt, 1995; Pitt & Samuel, 1993); in certain circumstances the latter approach could be more sensitive to subtle variation in a contextual effect than the ogive-based analysis. Accordingly, we conducted an ANOVA (with visual rate and perceived place of articulation as factors) on the percentage of voiceless responses in the ambiguous region for the fast and slow conditions for McGurk and non-McGurk responses. The outcome was highly similar to the ANOVA on the boundaries determined by the ogive-based analysis, with the only difference being that the marginal interaction of visual rate and perceived place of articulation reached significance ($p = .03$). All of the significant effects in the ogive-based analysis remained significant in the new analysis.

4. We note that the visual rate effect in Experiment 2, for both McGurk (1.94 msec) and non-McGurk (.96 msec) responses, was smaller in magnitude than the visual rate effect in Experiment 1 for audiovisually congruent stimuli (2.60 and 3.64 msec for the bilabial and alveolar conditions, respectively). It is possible that the audiovisual discrepancy in place of articulation in Experiment 2 caused a reduction in the effectiveness of the visual rate manipulation, but the difference across the experiments is difficult to interpret because it could also be due to the other methodological differences between Experiments 1 and 2 that we noted earlier.

(Manuscript received December 5, 2003;
revision accepted for publication September 22, 2004.)