Perception & Psychophysics
1989, 45 (1), 34-42

The role of visual information in the processing of place and manner features in speech perception

KERRY P. GREEN and PATRICIA K. KUHL
University of Washington, Seattle, Washington

Perception & Psychophysics
1989, 45 (1), 34-42

The role of visual information in the processing of place and manner features in speech perception

KERRY P. GREEN and PATRICIA K. KUHL
University of Washington, Seattle, Washington

Visual information provided by a talker's mouth movements can influence the perception of certain speech features. Thus, the "McGurk effect" shows that when the syllable /bi/ is presented audibly, in synchrony with the syllable /gi/, as it is presented visually, a person perceives the talker as saying /di/. Moreover, studies have shown that interactions occur between place and voicing features in phonetic perception, when information is presented audibly. In our first experiment, we asked whether feature interactions occur when place information is specified by a combination of auditory and visual information. Members of an auditory continuum ranging from /ibi/ to /ipi/ were paired with a video display of a talker saying /igi/. The auditory tokens were heard as ranging from /ibi/ to /ipi/, but the auditory-visual tokens were perceived as ranging from /idi/ to /iti/. The results demonstrated that the voicing boundary for the auditory-visual tokens was located at a significantly longer VOT value than the voicing boundary for the auditory continuum presented without the visual information. These results demonstrate that place-voice interactions are not limited to situations in which place information is specified audibly. In three follow-up experiments, we show that (1) the voicing boundary is not shifted in the absence of a change in the global percept, even when discrepant auditory-visual information is presented; (2) the number of response alternatives provided for the subjects does not affect the categorization or the VOT boundary of the auditory-visual stimuli; and (3) the original effect of a VOT boundary shift is not replicated when subjects are forced by instruction to "relabel" the /b-p/ auditory stimuli as /d/ or /t/. The subjects successfully relabeled the stimuli, but no shift in the VOT boundary was observed.

Early research on the perception of speech concentrated on identifying aspects of the acoustic signal that provide important "cues" for phonetic perception (Liberman, Delattre, & Cooper, 1958; Lisker & Abramson, 1964). This research demonstrated that rather than there being a single critical feature that governs the perception of a phonetic unit, many different acoustic events influence its perceived identity. More recently, research has focused on how these different acoustic cues are integrated during perception: many studies have demonstrated the occurrence of cue interactions (Fitch, Halwes, Erickson, & Liberman, 1980; Massaro & Oden 1980; Miller, 1977, 1981; Repp, 1983; Sawusch & Pisoni, 1974). Lisker and Abramson (1970) have shown, for example, that the location of the voiced/voiceless boundary changes along a VOT continuum as a function of the place of articulation of the phonetic segment. Thus, the voicing boundary on a bilabial continuum varying from /b/ to /p/ will occur at a shorter VOT value than the boundary on an alveolar /d-t/ continuum, which in turn will occur at a shorter VOT than a velar /g-k/ continuum.

Recent research on the auditory-visual perception of speech (see Summerfield, 1986 for review) has shown that phonetic identity is not only determined by auditory cues, but by visual ones as well. McGurk and MacDonald (1976; also MacDonald & McGurk, 1978), for example, have shown that the perceived place of articulation of a well-specified syllable such as /ba/ becomes changed by the presentation, in synchrony, of a video display of the speaker uttering a different syllable, such as /ga/. In this situation, observers typically perceive the speaker to be saying /da/. Several studies have since replicated and extended McGurk and MacDonald's original findings (Green & Kuhl, 1988; Kuhl, Green, & Meltzoff, 1988; Manuel, Repp, Liberman, & Studdert-Kennedy, 1983; Massaro & Cohen, 1983; Mills & Thiem, 1980; Summerfield, 1979).

The "McGurk effect" raises an interesting question regarding feature interaction: in studies that involve auditory information only, the processing of voicing interacts with the place of articulation of the phonetic segment; what happens when the place of articulation is specified by not only auditory but also visual information? Will the interaction between place and voicing still occur, or is the interaction limited to situations in which place information is specified solely by the auditory domain? The answer to this question contains implications for theories

This research was supported by NIH Grant HD-18286 to Patricia K. Kuhl. Kerry P. Green was supported by NIH Training Grant HD-07239 to the University of Washington. We wish to thank Karen Wolak for help in creating the stimuli, and Bruce Rindler and Julia McCormick for assistance in collecting the data. Part of the data was presented at the fall meeting of the Acoustical Society of America, Anaheim, CA, 1986. Inquiries should be sent to Kerry P. Green at the Child Development and Mental Retardation Center, WJ-10, University of Washington, Seattle, WA 98195.

about the interaction of place and voicing information. If the information that underlies voicing and place of articulation interacts only at a psychoacoustic level, then the VOT boundary should not shift when place information is presented visually. Alternatively, if other higher-order mechanisms are involved, then the interaction may still occur.

Roberts and Summerfield (1981) used the McGurk effect in order to ask whether the perceptual adaptation effect for speech (Eimas & Corbit, 1973) occurred at solely an auditory level or at a level that integrated information from both auditory and visual sources. Their results showed that adaptation was solely determined by the signal presented audibly, rather than by the phonetic percept, which was influenced by both auditory and visual information. Here we ask a similar question: Do the observed interactions between voicing and place occur only at the psychoacoustic level, or at a higher, more central level—one that integrates auditory and visual speech information?

In Experiment 1, auditory speech tokens drawn from an /ibi-ipi/ continuum were paired with a video display of the same speaker saying /igi/. When presented audibly, these tokens were all heard as /ibi/ or /ipi/. However, consistent with the McGurk effect (McGurk & MacDonald, 1976), when the stimuli were presented with visual /igi/, they were all perceived as having an alveolar place of articulation (/idi/ or /iti/). The specific question addressed in Experiment 1 was whether the location of the voiced/voiceless boundary along the VOT continuum was also influenced by this change in perceived place. If interactions between place and voicing are restricted to a psychoacoustic level, then the VOT boundary should be identical in the auditory and the auditory-visual conditions. If, however, the interaction takes place at a higher level, then there should be a shift in the VOT boundary towards longer VOT values in the auditory-visual condition in which the /ibi-ipi/ tokens are presented with visual /igi/ and are perceived to have an alveolar place.

EXPERIMENT 1

Method

Subjects

The subjects were 8 undergraduate students, who were given course credit as an incentive to participate. These subjects had no reported history of any speech or hearing disorder; all of them had normal or corrected-to-normal vision.

Materials

Visual stimuli. The visual stimuli were videotaped repetitions of a female speaker saying /igi/. A color camera (RCA TK-45), a microphone (Sony ECM-50), a special effects generator (Central Dynamics CDL 480-5), and a broadcast quality 1-in. videotape recorder (Sony BVH 1100 Type C) were used. The speaker was seated on a stool behind a piece of black velvet with a hole cut in it, which allowed only the speaker's face to be displayed on the camera. The camera was centrally focused on the speaker's lips. By means of the special-effects generator, the upper portion of the speaker's face (from the bottom of the nose to the top of the head) was replaced with video black. Only the region surrounding the speaker's mouth (including the jaw, lips, and oral cavity) was visible on videotape. Lighting of the face and the interior of the mouth was provided by two quartz lights (3,200°), one placed on either side of the speaker. This resulted in an excellent close-up view of the speaker's mouth region and of the interior of the oral cavity. Good resolution of the tongue position was provided during articulation.

A single /igi/ token was selected for the quality of the articulation and lack of extraneous movements or features. Five blocks of 13 repetitions of this single /igi/ were then created with a video editing console (JVC VE-92), connected to two ¾-in. videocassette machines (JVC CR8250). The repetitions were separated by approximately 1,300 msec of video black. Each repetition included 1 sec of video display of the speaker's face before the onset of articulation, the articulation of the /igi/, and 1 sec of the speaker's face after the offset of the /igi/ articulation. Each repetition was edited with an additional 1-sec fade-up from video black at the start of each repetition, and an additional 1-sec fade-out to video black at the end of the repetition. This prevented abrupt visual onsets or offsets of the trials, which could have caused masking or interference. The intertrial interval between the onset of 1 repetition and the onset of the next repetition was 6 sec.

Auditory stimuli. The auditory stimuli consisted of a voiced/voiceless series ranging from /ibi/ to /ipi/ and varying in VOT from 5 to 82 msec. The stimuli were computer-edited tokens of natural speech. They were created by recording a female speaker in a soundproof room, with an Electro-Voice microphone (No. 635A) and a Nagra III tape recorder. The syllables were digitized on a lab computer (LSI 11-73) at a 20 kHz sampling rate with a 9.89 kHz low-pass filter at 12 bits of amplitude quantization, and then analyzed using a signal-processing package. One /ibi/ and one /ipi/ with similar durations, both of which closely matched the durational characteristics of the visual /igi/ used to make the videotape, were selected for further processing. The stimuli were matched on the duration from the onset of the first vowel to consonantal release. A cross-splicing technique (see Ganong, 1980) was used to create a 13-member series that varied in VOT from 5 msec, the VOT of the original /ibi/ stimulus, to 82 msec, the VOT of the original /ipi/ stimulus. This was accomplished by deleting successively longer acoustic segments of the /ibi/, starting at the release of the consonant, and replacing them with acoustic segments of equal duration from the /ipi/, again starting at the release of the consonant. The durations of these acoustic segments—and therefore the VOT values of the stimuli—were 5, 17, 23, 28, 33, 39, 44, 49, 55, 60, 66, 71, and 82 msec. All cuts in /ibi/ were made at a zero-crossing at the beginning of a pitch period. There were no audible indications (e.g., the presence of a click) that any of the syllables had been edited; all sounded like good tokens of natural speech.

Auditory-visual stimuli. The auditory-visual stimuli were created by pairing each member of the /ibi-ipi/ series with the visual /igi/. The auditory stimuli were dubbed onto the videotape by playing /igi/ on a videocassette recorder (JVC CR8250). The output of one channel from the videotape recorder was fed into a Schmitt trigger on the LSI computer containing the auditory stimuli. This channel of the videotape contained the original auditory /igi/ corresponding to the visual /igi/ tokens. A marker tone preceded each /igi/ utterance. When the Schmitt trigger sensed the onset of the marker tone, it triggered the computer to output one of the members of the /ibi-ipi/ series; this occurred after a delay period, which was precalculated so that the release burst of the dubbed auditory syllable would precisely align with the release burst of the original /igi/ utterance. The syllables were output at a 20 kHz sampling rate, in a predetermined, randomized order; low-passed filtered at

9.89 kHz; and then recorded onto the second channel of the videotape. This procedure enabled us to dub the syllables onto the videotape with a high degree of accuracy.

Each of the 5 blocks of stimuli was copied twice in random order, with the stipulation that the same block could not occur twice in succession. The final test videotape contained a total of 10 blocks of 13 trials each, for a total of 130 trials. There was a break of approximately 10 sec between each block on the tape.

Procedure

Each subject participated in two half-hour sessions, conducted on different days. One session was an auditory-visual (AV) session, and the other, an auditory only (AO) session. The order of the test sessions was counterbalanced across subjects. In the AV session, the subjects were instructed to watch and listen to each trial, and to identify the stimuli as /idi/ or /iti/. In the AO session, the subjects were asked to identify the stimuli as /ibi/ or /ipi/. The subjects responded verbally in the AV condition by saying either /idi/ or /iti/ aloud to the experimenter, who recorded their responses on an answer sheet. This enabled the subjects to keep their attention and vision focused on the video monitor. During the AO session, the subjects responded by writing their responses on an answer sheet. Each subject was presented with a total of 260 trials—130 trials in the AV session, and another 130 trials in the AO session. This yielded 10 responses per subject for each stimulus in each of the AV or AO sessions. At the start of each session, the subjects were presented with one block of 13 practice trials. Their responses to these trials were not included in the analysis.

The subjects were tested individually in a small, dimly lit, sound-attenuated room. Each subject sat at a small desk located approximately 46 in. from a color video monitor (NEC JC-1215MA). At that distance, the open mouth of the talker displayed on the video monitor subtended a visual angle of 3.65° for the subject. The monitor was seated on a table behind a paper panel with a hole that allowed the subject to see it. The videotape was played on a videocassette player (JVC CP5550) located in an adjoining control room. During the AV session, the audio and video outputs from the videocassette player were presented via the video monitor. During the AO session, the video signal was disconnected; only the audio signal was presented via the loudspeaker in the video monitor. The contrast and brightness controls were both set at about midlevel, and the audio signal was presented at a comfortable listening level of approximately 65 dB SPL (A scale, fast), measured at the peak intensity of the second vowel, on a sound-level meter (Brüel & Kjaer 2203) placed at the appropriate distance and height of the subjects' heads.

Results and Discussion

The identification results for the AV and AO conditions are shown in Figure 1. In the identification function for the AV condition, a shift towards longer VOT values can be seen, relative to values for the AO function. The mean VOT boundaries for the AV and AO conditions provide a measure of the difference between the two functions. The individual voicing boundaries were determined by fitting a linear regression line to the data in the boundary region of each individual identification function, taking as the category boundary the VOT value that corresponded to 50% voiced responses. The mean boundaries were 49.25 and 43.12 msec for the AV and AO conditions, respectively.[1] A correlated t test on the difference between these two means was highly reliable [$t(7) = 9.58$, $p < .001$]. Moreover, all 8 subjects produced the shift in their VOT boundaries, and all the shifts were in the same direction.

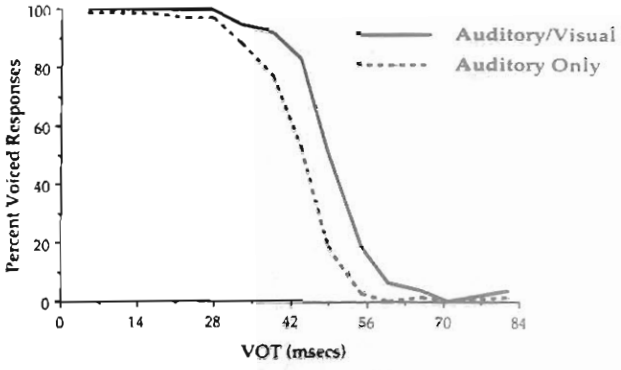

Figure 1. Percentage of voiced responses for the /ibi-ipi/ continuum under the auditory only and the auditory-visual conditions (Experiment 1).

The results of Experiment 1 suggest that the shift in the VOT boundary was due to the change in the perceived place of articulation brought about by the visual information. Three additional experiments were conducted to test alternative explanations for the findings. In Experiment 2, we tested the possibility that the shift in the VOT boundary was simply the result of presenting conflicting auditory-visual speech information. If conflicting visual information causes the shift, then whenever there is auditory-visual conflict, the VOT boundary should shift. In Experiment 3, we tested whether the shift in the VOT was caused by the number of response alternatives provided in the AO and the AV conditions. In Experiment 4, we tested whether asking subjects to simply relabel the auditory /ibi-ipi/ stimuli as /idi-iti/ would produce a shift in the VOT boundary.

EXPERIMENT 2

In order to address the conflicting visual information hypothesis, we required a situation in which the auditory and visual information conflicted without resulting in a change in the perceived place of articulation for the AV tokens. Previous research by MacDonald and McGurk (1978) had shown that when auditory /da/ is presented with visual /ga/, observers continue to report /da/. Therefore, this experiment involved pairing auditory tokens drawn from an /idi-iti/ continuum with the same visual /igi/ token as was used in Experiment 1.

We established that the visual /igi/ was identifiable as /igi/, and thus that it actually did provide conflicting visual information with the /idi-iti/ tokens, by conducting a visual-only experiment. Visual /igi/, /ibi/, and /idi/ were randomly presented 15 times each to a group of 5 subjects for identification. All stimuli were identified at better than 94% accuracy, with the visual /igi/ identified correctly 94.6% of the time. Therefore, the visual stimulus used in the present study was well-specified as /igi/ and

represented a suitable stimulus for providing conflicting visual information with the auditory /idi-iti/ stimuli.

Method

Subjects

The subjects were 8 new undergraduate students, who were given course credit as an incentive to participate. None reported any history of a speech or hearing disorder; all had normal or corrected-to-normal vision.

Materials

Visual stimuli. The visual stimuli consisted of the same five blocks of 13 repetitions of the single /igi/ that were created in Experiment 1.

Auditory stimuli. The auditory stimuli consisted of a voiced/voiceless series ranging from /idi/ to /iti/ and varying in VOT from 12 to 82 msec. These stimuli were also computer-edited tokens of natural speech and were created in the exact same fashion as the /ibi-ipi/ stimuli used in Experiment 1. The same female speaker was recorded as she produced tokens of /idi/ and /iti/. A single /idi/ token and a single /iti/ token, whose durational characteristics were similar to the /ibi/ and /ipi/ stimuli used in Experiment 1, and which closely matched the durations of the visual /igi/ used to make the videotape, were selected for further processing. The cross-splicing technique was used to create a 13-member series that varied in VOT from the 12 msec of the original /idi/ to the 82 msec of the original /iti/. The durations of the acoustic segments in this series were 12, 21, 27, 32, 35, 40, 45, 50, 56, 61, 66, 72, and 82 msec. These durations were selected because they closely matched the durations of the acoustic segments of the /ibi-ipi/ series in Experiment 1.

Auditory-visual stimuli. A single set consisting of five blocks of 13 trials of auditory-visual stimuli was created by means of the procedure described in Experiment 1, using the /idi-iti/ auditory stimuli.

Procedure

The procedure was almost the same as that used in Experiment 1, the exception being that the subjects were instructed to label the stimuli as either /idi/ or /iti/ in both the AO and AV conditions.

Results and Discussion

The identification results for the AV and AO conditions are presented in Figure 2, where it can be seen that there is essentially no difference in the identification functions for the two conditions. The mean VOT boundaries for the AV and the AO conditions were 51.4 and 52.0 msec respectively. A correlated t test indicated no significant difference between these two conditions ($t = -.52$, $p > .1$).

These results demonstrate that simply providing visual information that conflicts with the place of articulation of the auditory tokens is not enough to produce a shift in the VOT boundary. In Experiment 2, the same visual /igi/ as was used in Experiment 1, which was clearly discriminable from /idi/, did not produce a shift in the VOT boundary when it was paired with the /idi-iti/ auditory tokens. Thus, the perception of VOT appears to be influenced only when there is a change in the perceived place of articulation.

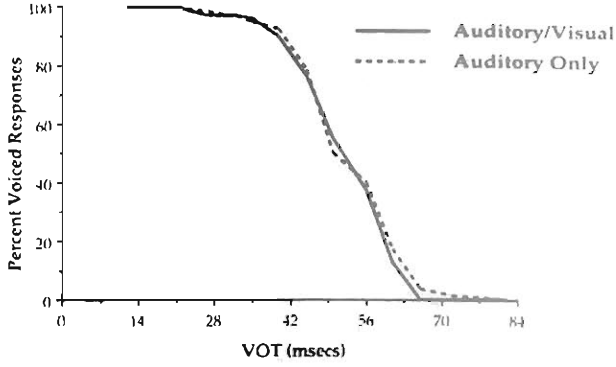

Figure 2. Percentage of voiced responses for the /idi-iti/ continuum under the auditory only and the auditory-visual conditions. (Experiment 2).

EXPERIMENT 3

In Experiment 3, we examined the effect of the number of response alternatives provided to the subjects. In Experiment 1, the subjects were asked to label the stimuli as /idi/ or /iti/. They were not given the option of using /ibi/ or /ipi/. This approach was taken because previous pilot work had demonstrated that the perception of /d/ was strong in this situation. Nevertheless, allowing the subjects to use all four response alternatives was considered important, because some of the subjects in Experiment 1 might not have perceived the auditory-visual stimuli as /idi/ and /iti/. In order to address this issue, and its effect on the boundary shift, a new group of subjects was presented with the AV stimuli of Experiment 1 under similar conditions. However, these new subjects were allowed to label the stimuli as /ibi/, /ipi/, /idi/, or /iti/.

A second group of subjects was presented with the AV stimuli from Experiment 2 under similar conditions. These subjects were allowed to label the stimuli as /idi/, /iti/, /igi/, or /iki/. The purpose of this condition was to determine whether the number of response alternatives changed the outcome of that experiment.

Method

Subjects

Twenty new subjects participated in this experiment. All received course credit for their participation; none reported any history of a speech or hearing disorder. all had normal or corrected-to-normal vision.

Stimuli

The stimuli were the same as those used in the AV conditions for Experiments 1 and 2.

Procedure

Half of the subjects were presented with the AV stimuli from Experiment 1, and the other half were presented with the AV stimuli from Experiment 2. The procedures were identical to those used in the AV conditions of Experiments 1 and 2, with the following

exceptions: the subjects presented with the stimuli from Experiment 1 were asked to identify the stimuli as either /ibi/, /ipi/, /idi/, or /iti/, and the subjects presented with the stimuli from Experiment 2 were asked to identify the stimuli as either /idi/, /iti/, /igi/, or /iki/.

Results and Discussion

The percent responses in each of the four categories for the two experimental conditions are presented in Table 1. Consider first the results for the /ibi-ipi/ stimuli used in Experiment 1. The results demonstrate that when these auditory stimuli are presented with the visual /igi/, they are categorized as /idi/ or /iti/ over 96% of the time. The individual VOT boundaries for each subject were also calculated, using the procedure described in Experiment 1. The mean VOT boundary for this four-choice condition (48.1) is very similar to the two-choice AV condition in Experiment 1 (49.25). An independent t test confirmed that there was no reliable difference between the two conditions ($t = .51$, $p > .1$). There was, however, a reliable difference between the four-choice condition and the AO condition (43.2) of Experiment 1 ($t = 2.07$, $p < .05$). Therefore, the shift in the VOT boundary between the AO and AV conditions of Experiment 1 cannot be attributed solely to the restricted use of the labels /idi/ and /iti/ in the AV condition.

The results for the /idi-iti/ stimuli are also shown in Table 1. These stimuli were categorized as /idi/ or /iti/ over 92% of the time. As suggested by Experiment 2, and by previous reports in the literature, the conflicting visual /igi/ stimulus had very little impact on the categorization of these stimuli, even though the visual stimulus was clearly identifiable as /igi/. In addition, there was no effect of the visual information on the VOT boundary. The mean VOT for this four-choice condition (53.62) was very similar to the mean VOT boundaries of both the AO (52.0) and AV (51.37) conditions of Experiment 2 (both $ts < .7$; $p > .1$).

The results of Experiment 3 thus indicate that when subjects are allowed to use an expanded set of phonetic labels in the AV condition, there is still a reliable shift in the VOT boundary for the situation in which the auditory /ibi-ipi/ stimuli are combined with /igi/, but not when the auditory /idi-iti/ are combined with /igi/. In other words, the effects of Experiments 1 and 2 were not altered by the number of response alternatives provided to the subjects.

Table 1
Percent Responses for Each Category in Experiment 1

Continuum of Auditory Stimuli	Response Choices					
	/b/	/p/	/d/	/t/	/g/	/k/
/ibi-ipi/	1.1	2.1	54.9	41.9	—	—
/idi-iti/	—	—	56.1	36.8	6.1	1.0

Note—Blanks = response choice not applicable for that stimulus continuum.

EXPERIMENT 4

Research by Carden, Levitt, Jusczyk, and Walley (1981) has demonstrated that asking subjects to relabel a phonetic /ba-da/ continuum as either /fa/ or /θa/ can shift the phonetic boundary of the place of articulation. According to Carden et al., the relabeling instructions themselves were presumed to cause the subjects to perceive the auditory tokens as having a different manner of articulation, probably because their synthetic speech stimuli were ambiguous with regard to the manner of articulation.

The issue raised by the findings of Carden et al. (1981) with regard to the present experiment is whether the effect of the visual information in Experiment 1 could be mimicked by simply asking subjects to relabel the auditory /ibi-ipi/ stimuli as /idi/ or /iti/. We predicted that the auditory stimuli used in these experiments were unambiguous, and that the place information was specified so well that asking observers to relabel them would have no effect on their perceived place of articulation, and therefore no effect on the VOT boundary. Such an outcome would indicate that the shift in the VOT boundary observed in Experiment 1 could not be mimicked by a simple relabeling of the auditory stimuli.

Method

Subjects

Thirty-two new subjects participated in the experiment. None reported any history of a speech or hearing disorder. They received course credit for their participation.

Materials

The stimuli consisted of the same /ibi-ipi/ and /idi-iti/ auditory tokens as were used in Experiments 1 and 2.

Procedure

Different groups of 8 subjects were run on each of the four different experimental conditions. Each subject was presented with either the /ibi-ipi/ or the /idi-iti/ stimuli. They were instructed to label the /ibi-ipi/ stimuli as /ibi/ or /ipi/ in one condition and as /idi/ or /iti/ in a second condition. In a third and a fourth condition, the subjects were instructed to label the /idi-iti/ stimuli as (1) /idi/; (2) /iti/ or /ibi/; or (3) /ipi/, respectively. The four groups of subjects were each tested during a single ½-h session. The subjects were instructed to respond with only the labels provided; they wrote their responses on an answer sheet. At the start of the session, each subject was presented with a block of practice trials. The responses to the practice trials were not included in the analysis. The subjects were tested with the same apparatus as was used for the AO conditions in Experiments 1 and 2.

Results and Discussion

As predicted, none of the subjects reported a change in the perceived place of articulation of the auditory stimuli as a result of the labeling instructions. They continued to hear the bilabial stimuli as /ibi/ and /ipi/, and the alveolar stimuli as /idi/ and /iti/. Nonetheless, the subjects

relabeled the stimuli with the labels they were instructed to use, and all the subjects produced consistent labeling functions. The subjects seemed to perform the relabeling task by using a rule in which, for example, if they heard /ibi/, they would label the stimulus /idi/, and if they heard /ipi/, they would label it /iti/—even though no such strategy was suggested during the instructions. Apparently, asking the subjects to label the auditory stimuli with a different place of articulation had no influence on their perceptual experience; they perceived the same place of articulation regardless of the instructions. What is particularly interesting is that the visual information influenced the subjects' perception of the place of articulation and their processing of VOT information even when the auditory stimuli were clear and unambiguous, whereas relabeling did not.

Since the relabeling instructions did not affect the perceived place of articulation of the auditory stimuli, one would not expect the instructions to have had any effect on the VOT boundaries. The labeling functions for the /ibi-ipi/ and /idi-iti/ stimuli are presented in Figure 3. The individual VOT boundaries were calculated for each subject in each experimental condition, using the procedure described in Experiment 1. These boundaries were analyzed by means of a two-factor analysis of variance, with auditory stimulus and labeling condition as the two factors. The analysis indicated that only the effect of auditory stimulus was significant [$F(1,28) = 19.7$, $p < .0005$]. The effect of labeling [$F(1,28) = .22$, $p > .6$], and the interaction between the two effects [$F(1,28) = .42$, $p > .5$], did not approach significance. Furthermore, the boundaries for the /ibi-ipi/ stimuli (39.1 and 41.4, labeled as ibi/ipi and idi/iti, respectively) are in good agreement with the boundaries for the same auditory stimuli under the AO condition in Experiment 1. Moreover, the boundaries for the /idi-iti/ stimuli (49.4 and 49.0, labeled as ibi/ipi and idi/iti, respectively) are similar to the boundaries for the same stimuli under the AO condition in Experiment 2.

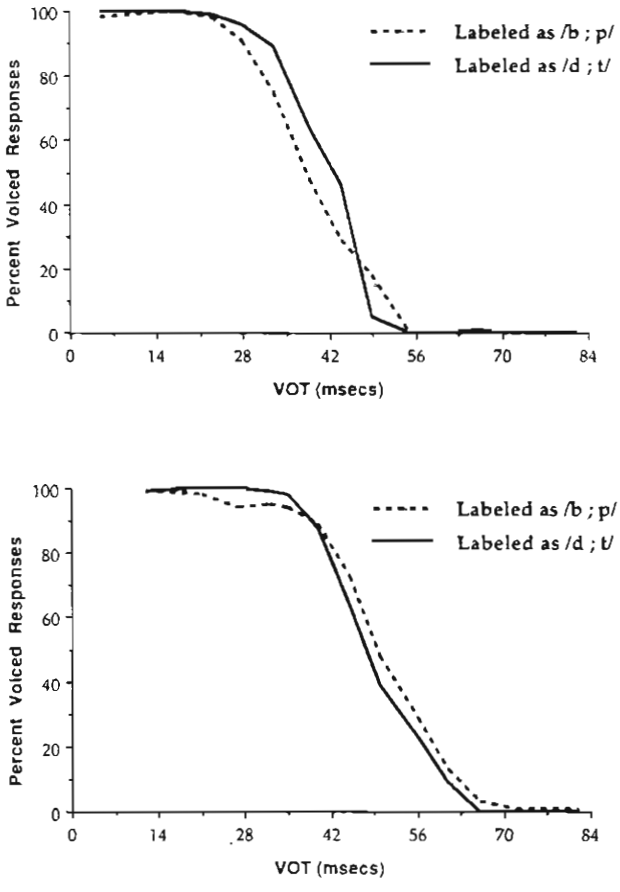

Figure 3. Percentage of voiced responses for the /ibi-ipi/ continuum (upper graph) and the /idi-iti/ continuum (lower graph) for the different (re)labeling conditions (Experiment 4).

In summary, there appears to be a difference, at least for our stimuli, between the feature interactions observed in a relabeling experiment and those produced as a result of an auditory-visual illusion. In the relabeling situations, the subjects did not report a percept shift; they simply gave the stimulus they heard a new label. Giving it a new label was not sufficient to shift the VOT boundary. Conversely, in the auditory-visual illusion situation, an unambiguous change in the percept occurred, one that was not phenomenally recognized, and this change in percept was accompanied by a change in the VOT boundary.

GENERAL DISCUSSION

The present investigation was motivated by two recent findings: (1) that auditory information about place of articulation interacts with the processing of voicing information during phonetic perception, and (2) that information from both the auditory and visual modalities contributes to the perception of place of articulation. The experiments in this investigation addressed the question of whether visual information influences not only the perceived place of articulation of phonetic segments, but also the processing of VOT information.

The approach taken to address this question was to pair auditory tokens drawn from either an /ibi-ipi/ or /idi-iti/ continuum with a visual /igi/ token, and then to assess whether the VOT boundary in the auditory-visual situation was different from the boundary for the auditory tokens alone. Our results demonstrate that the effect on the processing of VOT information is tied to a change in the perception of place of articulation. A change in the former only occurs when a change in the latter is somehow induced. The visual stimuli induced a change in perception of place, so a corresponding change in VOT was observed. Simple relabeling did not induce a perceived change in the place of articulation, and therefore no corresponding change in VOT was observed in that case.[3]

The results from this study contain implications for theories of the integration of auditory and visual information in speech perception. In addition, the results suggest important implications for models of phonetic perception,

specifically for models that account for the interaction of place and voicing information.

With regard to theories of the integration of auditory and visual information in speech perception, one important issue involves the level at which this integration occurs. That is, when, in the processing of the speech information, do the auditory and visual streams combine? When most auditory experiments on place-voicing interactions are conducted, place information is well-specified. The continua have spectral cues specifying either a bilabial, alveolar, or velar place of articulation, and this information can immediately interact with the voicing information. In the experiments reported here, information from both the auditory channel (which specifies /b/) and the visual channel (which specifies /g/) must be combined in order to determine the perceived place of articulation (/d/). These results raise the question of the order in which the information is integrated in auditory-visual speech perception.

Two possibilities present themselves. The first is that the auditory and visual place information are initially integrated, resulting in an alveolar place specification prior to the determination of voicing. This alveolar place specification then interacts with the perception of voicing, causing a shift in the VOT boundary. According to this approach, the VOT information is processed with regard to one of three boundaries depending upon the place of articulation. If the place is specified as bilabial, a relatively shorter boundary is used (around 43 msec in the present experiments). If the place is alveolar, a slightly longer boundary is used (approximately 49 msec in our experiments), and so on. This approach is similar to Carden et al.'s (1981) two-boundary hypothesis for the interaction of place and manner information. Now consider how such an approach accounts for the results of the current study. The visual and auditory place information would first be analyzed independently (see Massaro & Cohen, 1983) and then integrated to produce a place specification consistent with the information in the two modalities. This place specification would then be used to determine the particular boundary against which the VOT information is judged.

The second possibility is that all three sources of information—auditory place, visual place, and auditory voicing information—converge and interact at the same time. According to this approach, the observer perceives the sequence of phonemes that are most consistent with the combined auditory and visual information (Summerfield, 1986). This approach accounts for the present findings by claiming that an auditory token from the /ibi-ipi/ continuum paired with the visual /igi/ is consistent with /iti/ if the VOT exceeds 45 msec, and consistent with /idi/ if the VOT is less than 45 msec. The crucial difference between these two approaches is the order in which the visual and auditory place information is extracted and compared with the auditory voicing information. The first approach claims that the auditory-visual place information is integrated first, and that it then influences the evaluation of the auditory voicing information. The second approach claims that the auditory place, visual place, and auditory voicing information are processed simultaneously. We are currently conducting experiments to address this issue.

With regard to the second issue, the nature of feature interactions in phonetic perception, one approach has been to account for such interactions by arguing that they are psychoacoustic in nature. The evidence in support of this approach comes primarily from two types of experiments: studies showing that the interactions of place and voicing information observed in the speech perception of human listeners (Lisker & Abamson, 1970) are replicable in animals; and experiments on adults that involve complex nonspeech signals analogous to speech. The studies of animals' perception of speech demonstrate that both chinchillas and monkeys produce VOT boundaries similar to humans, across the different places of articulation for stop consonants (Kuhl & Miller, 1978; Kuhl & Padden, 1982). These findings, which indicate that a general auditory mechanism is available for the perception of VOT in animals, raise the strong possibility that such a mechanism also underlies the perception of voicing in humans (Kuhl, 1986, 1987b).

In the studies employing nonspeech analogues to VOT stimuli (e.g., Miller, Wier, Pastore, Kelly, & Dooling, 1976; Pisoni, 1977), it has been argued that the perception of voicing, which involves a judgment of the temporal order of certain auditory events (such as aspiration and the onset of voicing) is accomplished by a psychoacoustic mechanism responsible for the judgment of the temporal order of auditory events. According to this view, there is a psychoacoustic threshold of approximately 20 msec; if two auditory events are separated by less than this amount, they are perceived as occurring simultaneously, and if they are separated by more than 20 msec, they are perceived as successive (Hirsh, 1959; Stevens & Klatt, 1974).

Nevertheless, although the psychoacoustic threshold is close to the bilabial VOT boundary of 25 msec, it is very different from the 35-msec alveolar and the 42-msec velar boundaries reported by Lisker and Abramson (1970). The challenge for a psychoacoustic account is to explain why a psychoacoustic threshold should shift as a function of place of articulation (Kuhl, 1982). One possible answer lies in the fact that Lisker and Abramson created their different VOT continua with different onset frequencies of F1 for the different places of articulation. Other studies have demonstrated that the onset frequency of F1 infuences the VOT boundary (Lisker, 1975; Summerfield & Haggard, 1977). It is therefore possible that a psychoacoustic explanation underlies the interaction of F1 onset frequency and VOT, which would account for the shift in VOT across place of articulation. Yet it has been

difficult to demonstrate such an interaction by means of employing nonspeech analogue stimuli (Summerfield, 1982; but see Hillenbrand, 1984; Parker, 1988).

Also relevant are the findings of Miller (1977). She created auditory stimuli in which the only acoustic difference across the different place continua was the starting frequency of the second formant. All of the other acoustic aspects of the stimuli—such as the onset frequency of F1—were held constant. The results showed that the shifts in the boundary locations for the differing places of articulation remained, though they were reduced from the 10 msec for the bilabial to alveolar and alveolar to velar places seen in the original Lisker and Abramson (1970) study, to 2–3 msec. In other words, equalizing the F1 onset cues across place reduced the effects a great deal, and this presumably did account for some portion of the orginally observed shift in the location of the boundary. Nevertheless, the much smaller shifts in the location of the boundary that remained were significant. A psychoacoustic account of these results would require an explanation of how the onset frequencies of F2 and F3 influence the threshold for temporal order judgments; thus far, such an explanation has not been provided (see Kuhl, 1982).

Miller's (1977) findings raise the possibility that not all the interactions between place and voicing can be accounted for solely by means of a psychoacoustic mechanism. The data from the present study contribute to this argument by demonstrating that in adult humans, the place–voicing interaction is not restricted to situations in which the information is provided solely through the auditory channel; visual information that changes the perceived place of articulation has the same effect on the perception of voicing. The VOT boundary was affected even though the auditory stimuli were unchanged between the AO and AV conditions of Experiment 1, which rules out a purely psychoacoustic explanation. The findings from the current investigation and from other studies on the influence of visual information on phonetic perception (see Summerfield, 1986 for review) appear to demonstrate that adult humans are capable of taking more than auditory information into account when processing feature information, and that therefore they must have access to mechanisms beyond those that are psychoacoustic in nature.

A related issue is the development of such abilities. There are some indications that infants process VOT information in relation to F1 onset frequency in a manner similar to adult listeners (Miller & Eimas, 1983). However, whether infants are relying on psychoacoustic mechanisms for the perception of voicing or on more complex mechanisms is an open question (Kuhl, 1987a; Liberman, 1982). Jusczyk (1985) has proposed that infants rely on psychoacoustic mechanisms initially, and develop a phonetic mechanism for perceiving speech only later on, as they acquire language. Studies have already shown that young infants can detect the correspondence between audibly present speech sound and the sight of a person producing that sound (Kuhl & Meltzoff, 1982). What is not known is whether young infants integrate conflicting auditory and visual place information, and, therefore, whether featural interactions like those observed in this study would occur in infants. The answer to this question will have important implications for theories of developmental speech perception; but regardless of the outcome of such developmental studies, the work presented here demonstrates that in adults, featural interactions in speech take into account information from both the auditory and the visual modalities.

REFERENCES

Carden, G., Levitt, A., Jusczyk, P. W., & Walley, A. (1981). Evidence for phonetic processing of cues to place of articulation: Perceived manner affects perceived place. *Perception & Psychophysics*, **29**, 26-36.

Eimas, P. D., & Corbit, J. D. (1973). Selective adaptation of linguistic feature detectors. *Cognitive Psychology*, **4**, 99-109.

Fitch, H. L., Halwes, T., Erickson, D. M., & Liberman, A. M. (1980). Perceptual equivalence of two acoustic cues for stop-consonant manner. *Perception & Psychophysics*, **27**, 343-350.

Foster, G. A. (1984). Where do features interact? *Proceedings of the Institute of Acoustics*, **6**, 393-400.

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception & Performance*, **6**, 110-125.

Green, K. P., & Kuhl, P. K. (1988). The interaction of visual place and auditory voicing information during the perception of speech. *Journal of the Acoustical Society of America*, **83**, S85.

Hillenbrand, J. (1984). Perception of sine-wave analogs of voice onset time stimuli. *Journal of the Acoustical Society of America*, **75**, 231-240.

Hirsh, I. J. (1959). Auditory perception of temporal order. *Journal of the Acoustical Society of America*, **31**, 759-767.

Jusczyk, P. (1985). On characterizing the development of speech perception. In J. Mehler & R. Fox (Eds.), *Neonate cognition: Beyond the blooming buzzing confusion* (pp. 199-229). Hillsdale, NJ: Erlbaum.

Kuhl, P. K. (1982). Speech perception: An overview of current issues. In N. J. Lass, L. V. McReynolds, J. L. Northern, & D. E. Yoder (Eds.), *Speech, language, and hearing: Vol. 1. Normal processes* (pp. 286-322). Philadelphia, PA: W. B. Saunders.

Kuhl, P. K. (1986). Theoretical contributions of tests on animals to the special-mechanisms debate in speech. *Experimental Biology*, **45**, 233-265.

Kuhl, P. K. (1987a). Perception of speech and sound in early infancy. In P. Salapatek & L. Cohen (Eds.), *Handbook of infant perception: Vol. II. From perception to cognition* (pp. 275-382). New York: Academic Press.

Kuhl, P. K. (1987b). The special-mechanisms debate in speech research: Categorization tests on animals and infants. In S. Harnad (Ed.), *Categorical perception: The groundwork of cognition* (pp. 355-386). Cambridge, England: Cambridge University Press.

Kuhl P. K., Green, K. P., & Meltzoff, A. N. (1988). Factors affecting the integration of auditory and visual information in speech: The level effect. *Journal of the Acoustical Society of America*, **83**, S86.

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

Kuhl, P. K., & Miller, J. D. (1978). Speech perception by the chinchilla: Identification functions for synthetic VOT stimuli. *Journal of the Acoustical Society of America*, **63**, 905-917.

Kuhl, P. K., & Padden, D. M. (1982). Enhanced discriminability at the phonetic boundaries for the voicing feature in macaques. *Perception & Psychophysics*, **32**, 542-550.

Liberman, A. M. (1982). On the finding that speech is special. *American Psychologist*, **37**, 148-167.

LIBERMAN, A. M., DELATTRE, P., & COOPER, F. S. (1958). Distinction between voiced and voiceless stops. *Language & Speech*, **1**, 153-167.

LISKER, L. (1975). Is it VOT or a first-formant detector? *Journal of the Acoustical Society of America*, **57**, 1547-1551.

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., & 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). Exploring the "McGurk Effect." *Journal of the Acoustical Society of America*, **74**, S66.

MASSARO, D., & COHEN, 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., & ODEN, G. C. (1980). Evaluation and integration of acoustic features in speech perception. *Journal of the Acoustical Society of America*, **67**, 996-1013.

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

MILLER, J. D., WIER, C. C., PASTORE, R. E., KELLY, W. J., & DOOLING, R. J. (1976). Discrimination and labeling of noise-buzz sequences with varying noise-lead times: An example of categorical perception. *Journal of the Acoustical Society of America*, **60**, 410-417.

MILLER, J. L. (1977). Nonindependence of feature processing in initial consonants. *Journal of Speech & Hearing Research*, **20**, 510-518.

MILLER, J. L. (1981). Phonetic perception: Evidence for context-dependent and context-independent processing. *Journal of the Acoustical Society of America*, **69**, 822-831.

MILLER, J. L., & EIMAS, P. D. (1983). Studies on the categorization of speech by infants. *Cognition*, **13**, 135-165.

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

PARKER, E. M. (1988). Auditory constraints on the perception of voice-onset time: The influence of lower tone frequency on judgments of tone-onset simultaneity. *Journal of the Acoustical Society of America*, **83**, 1597-1607.

PISONI, D. B. (1977). Identification and discrimination of the relative onset time of two component tones: Implications for voicing perception in stops. *Journal of the Acoustical Society of America*, **61**, 1352-1361.

REPP, B. H. (1983). Trading relations among acoustic cues in speech perception are largely a result of phonetic categorization. *Speech Communication*, **2**, 341-361.

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

SAWUSCH, J. R., & PISONI, D. B. (1974). On the identification of place and voicing features in synthetic stop consonants. *Journal of Phonetics*, **2**, 181-194.

STEVENS, K. N., & KLATT, D. H. (1974). Role of formant transitions in the voiced-voiceless distinction for stops. *Journal of the Acoustical Society of America*, **55**, 653-659.

SUMMERFIELD, Q. (1979). Use of visual information for phonetic perception. *Phonetica*, **36**, 314-331.

SUMMERFIELD, Q. (1982). Differences between spectral dependencies in auditory and phonetic temporal processing: Relevance to the perception of voicing in initial stops. *Journal of the Acoustical Society of America*, **72**, 51-61.

SUMMERFIELD, Q. (1986). Some preliminaries to a comprehensive account of audio-visual speech perception. In R. Campbell & B. Dodd (Eds.), *Hearing by eye* (pp. 3-51). London: Erlbaum.

SUMMERFIELD, Q., & HAGGARD, M. P. (1977). On the dissociation of spectral and temporal cues to the voicing distinction in initial stop consonants. *Journal of the Acoustical Society of America*, **62**, 435-448.

NOTES

1. These VOT boundaries are longer than are typically found for computer-synthesized bilabial stop consonants, probably because the continuum consisted of edited tokens of natural speech in which a number of different acoustic cues for voicing were present. The exact location of the boundary does not bear on the major finding of this experiment, which is that presenting the same auditory information in conjunction with the visual information causes the VOT boundary to shift towards reliably longer VOT values.

2. We are grateful to a reviewer for bringing to our attention a study by Foster (1984) in which a similar finding was described.

(Manuscript received August 20, 1987;
revision accepted for publication July 18, 1988.)