Perception & Psychophysics
2003, 65 (4), 591-601

Visual influences on the internal structure of phonetic categories

LAWRENCE BRANCAZIO, JOANNE L. MILLER, and MATTHEW A. PARÉ
Northeastern University, Boston, Massachusetts

Previous work has demonstrated that the graded internal structure of phonetic categories is sensitive to a variety of contextual factors. One such factor is place of articulation: The best exemplars of voiceless stop consonants along auditory bilabial and velar voice onset time (VOT) continua occur over different ranges of VOTs (Volaitis & Miller, 1992). In the present study, we exploited the *McGurk effect* to examine whether visual information for place of articulation also shifts the best-exemplar range for voiceless consonants, following Green and Kuhl's (1989) demonstration of effects of visual place of articulation on the location of voicing boundaries. In Experiment 1, we established that /p/ and /t/ have different best-exemplar ranges along auditory bilabial and alveolar VOT continua. We then found, in Experiment 2, a similar shift in the best-exemplar range for /t/ relative to that for /p/ when there was a change in visual place of articulation, with auditory place of articulation held constant. These findings indicate that the perceptual mechanisms that determine internal phonetic category structure are sensitive to visual, as well as to auditory, information.

A traditional approach for investigating speech perception is to explore the perceptual boundaries between phonetic categories. Typically, examinations of phonetic boundaries involve a phoneme identification task, using speech sounds taken from a continuum varying along an acoustic dimension. One phonetic dimension that has been explored extensively with respect to perceptual boundaries is voicing. Voiced initial stop consonants (/b/, /d/, and /g/) differ from their voiceless counterparts (/p/, /t/, and /k/) primarily in voice onset time (VOT), the interval between consonant release and the onset of vocal-fold vibration, such that voiced consonants have shorter VOTs than do voiceless consonants. Accordingly, when listeners are presented with stimuli from a continuum of sounds varying only in VOT, they tend to identify the consonants with short VOTs as voiced and those with long VOTs as voiceless, with the perceptual boundary occurring at an intermediate VOT.

Studies in which phonetic categorization has been examined have consistently demonstrated that the locations of voiced–voiceless boundaries along a VOT continuum are highly sensitive to a variety of contextual factors (Repp & Liberman, 1987). One such factor is the acoustic-phonetic property, place of articulation. It is well established that perceptual voicing boundaries for bilabial (/b/ and /p/), alveolar (/d/ and /t/), and velar (/g/ and /k/) stop consonants occur at successively longer VOTs (Lisker & Abramson, 1970). This shift in boundaries is consistent with the differences in speech production in VOT across the different places of articulation, with successively longer VOTs for /p/, /t/, and /k/ (Lisker & Abramson, 1964). Another acoustic-phonetic factor that affects voicing boundaries is speaking rate. Voicing boundaries (e.g., between /b/ and /p/) occur at shorter VOTs for short syllables, which reflect a fast speaking rate, than for long syllables, which reflect a slow speaking rate (Summerfield, 1981). Voicing boundaries are also influenced by contextual factors that involve higher order knowledge, such as lexical status. Specifically, stimuli along a VOT continuum tend to be identified more often in such a way that they form words rather than nonwords, resulting in a boundary shift (e.g., Ganong, 1980; Miller & Dexter, 1988; Pitt, 1995). For example, when a final /f/ and a final /s/ are appended to the stimuli from an auditory /bi/–/pi/ continuum to create one series that varies from a word to a nonword (*beef–peef*) and another series that varies from a nonword to a word (*beace–peace*), the voicing boundary occurs at a shorter VOT along the *beace–peace* series than along the *beef– peef* series.

Thus, a variety of contextual factors have systematic effects on phonetic boundaries along auditory continua varying in VOT. Interestingly, contextual effects are not limited to the auditory domain; research has shown that voicing boundaries are also sensitive to *visually* specified contextual factors. One example was provided in a study by Green and Kuhl (1989) that examined whether visual place of articulation, like auditory place of articulation, has a systematic effect on voicing boundary locations. To

This research was supported by NIH Postdoctoral Fellowship F32 DC00373, awarded to the first author, and by NIH Grant R01 DC00130, awarded to the second author. Revision of the manuscript was partially supported by NIH Grant HD01994, awarded to Haskins Laboratories. The authors thank Sean Allen, Michèle Mondini, and two anonymous reviewers for helpful comments on earlier versions of this manuscript and Elizabeth Baraff and Sarah Whiting for testing the subjects. Correspondence concerning this article should be addressed to L. Brancazio, Department of Psychology, Southern Connecticut State University, 501 Crescent St., New Haven, CT 06515 (e-mail: brancazio@southernct.edu).

test this, Green and Kuhl exploited the *McGurk effect* (McGurk & MacDonald, 1976), a phenomenon that occurs when an auditory signal specifying a particular phonetic segment is presented in conjunction with a visual signal specifying a different phonetic segment with a distinct place of articulation (e.g., an auditory utterance of /ba/ presented with a visual /da/ or /ga/). When presented with these stimuli, subjects often report that they perceive a phonetic segment different from the one specified by the auditory signal (e.g., an auditory /ba/ with a visual /da/ or visual /ga/ is often identified as /da/), reflecting the perceptual integration of the information across the two modalities (MacDonald & McGurk, 1978; Massaro, 1987, 1998). Green and Kuhl presented auditory stimuli from an /ibi/–/ipi/ continuum in an auditory-only condition and in an audiovisual condition in which the auditory stimuli were presented with an incongruent visual token of /igi/. Because the incongruent visual stimulus gave rise to a McGurk effect, subjects perceived /idi/ and /iti/ in the audiovisual condition, whereas they perceived /ibi/ and /ipi/ in the auditory-only condition. Importantly, the voicing boundary between perceived /d/ and /t/ in the audiovisual condition occurred at a longer VOT than the boundary between /b/ and /p/ in the auditory-only condition. This shift parallels the difference in voicing boundaries typically found between auditory /b/–/p/ and /d/–/t/ continua, even though Green and Kuhl held constant the auditory properties of the stimuli relating to place of articulation. Green and Kuhl's findings indicate that with respect to voicing, perceivers use visual, as well as auditory, information about place of articulation when determining boundary locations.

The studies discussed thus far have examined the effects of various contextual factors on the perception of voicing by demonstrating shifts in the boundaries *between* voiced and voiceless consonants. However, researchers in recent years have also investigated listeners' perceptual sensitivity to contextual factors *within* the perceptual category for voiceless consonants. Considerable evidence now exists that phonetic categories have a well-defined graded internal structure, such that certain tokens within a category are considered better exemplars of the category than are other tokens (e.g., Kuhl, 1991; Miller, 1994; Samuel, 1982). Several studies have specifically addressed internal category structure with respect to VOT, using a goodness-rating task. Subjects are presented with stimuli from extended VOT continua, ranging from very short to extremely long VOTs, and are asked to rate each stimulus according to how good an exemplar of a particular phonetic category it is (Allen & Miller, 2001; Miller & Volaitis, 1989; Volaitis & Miller, 1992; Wayland, Miller, & Volaitis, 1994). Along such a continuum, stimuli with the shortest VOTs are perceived as voiced consonants (e.g., /b/), stimuli with somewhat longer VOTs are perceived as voiceless consonants (e.g., /p/), and stimuli with the longest VOTs are perceived as breathy, exaggerated versions of the voiceless consonant (labeled with *, as in */p/). Accordingly, when subjects are asked to rate stimuli from an extended VOT continuum as exemplars of a voiceless consonant (e.g., rating stimuli from a /b/–/p/–*/p/ continuum as exemplars of /p/), the stimuli falling within a certain range of VOTs receive very high ratings, and ratings systematically drop as VOT increases or decreases from this best-exemplar range. In these studies, the *best-exemplar range* for a particular continuum is determined by identifying the range of VOTs at which the ratings are higher than some criterial value (for example, 90% of the highest rating for the continuum) and by locating the lower and upper limits of this range.

In several studies, Miller and colleagues have demonstrated that the internal structure of phonetic categories can be systematically affected by contextual variables in a manner that is consistent with the variables' effects on category boundaries (e.g., Hodgson & Miller, 1996; Miller, 1994; Miller, O'Rourke, & Volaitis, 1997). Two factors that have been shown to affect the internal category structure of voiceless consonants are speaking rate and place of articulation. Miller and Volaitis (1989; see also Allen & Miller, 2001; Volaitis & Miller, 1992) found that the effect of speaking rate not only occurs at voicing boundaries (i.e., as was noted earlier, the voicing boundary for short syllables occurs at a shorter VOT than does the boundary for long syllables), but also extends throughout the best-exemplar range for /p/. Specifically, they found that both the lower and the upper limits of the best-exemplar range for /p/ occurred at shorter VOTs for a /bi/–/pi/–*/pi/ continuum with short stimuli than for a continuum with long stimuli (see also Wayland et al., 1994, for an effect of sentential speaking rate on best-exemplar ranges). Evidence for an effect of place of articulation on the internal category structure of voiceless consonants was provided by Volaitis and Miller, who found that the best exemplars of /ki/ along a /gi/–/ki/–*/ki/ continuum occurred over a range of longer VOTs than did the best exemplars of /pi/ along a /bi/–/pi/–*/pi/ continuum, again evidenced by shifts in both the lower and upper limits of the respective best-exemplar ranges. This shift is consistent with the finding, noted earlier, that the /g/–/k/ boundary occurs at a longer VOT than does the /b/–/p/ boundary (Lisker & Abramson, 1970). Together, these findings indicate that contextual variables can have effects that are not limited to the boundaries between voicing categories but that extend far into the voiceless consonant category.

Interestingly, though, recent work has demonstrated that not all contextual variables that affect voicing boundaries also affect internal category structure. In a recent study, Allen and Miller (2001) tested whether the effect of lexical status on category boundaries, described earlier, would extend throughout the voiceless category. They compared goodness ratings for /p/ in matched *beace–peace–*peace* and *beef–peef–*peef* continua (where /p/ yielded a word in the former continuum but a nonword in the latter) and found that lexical status affected only the portion of the best-exemplar range closest to the /b/–/p/ category boundary; that is, there was a shift in the lower limit, but not in the upper limit, of the best-exemplar range.

This outcome contrasts with the effects of speaking rate and place of articulation, which entail a shift in the entire best-exemplar range—that is, in both the lower and the upper limits of the range. Thus, with respect to the perception of voicing, the effect of lexical status, a higher order contextual factor, is much more limited in its extent than are the effects of the acoustic-phonetic factors of speaking rate and place of articulation.

The available evidence thus indicates that acoustic-phonetic contextual factors that affect voicing boundaries also affect the internal category structure of voiceless consonants but that this internal category structure is not sensitive to all contextual factors that affect boundaries. The purpose of the present study was to examine whether the internal category structure of voiceless consonants is sensitive to a visual contextual factor—namely, visual place of articulation. As we noted earlier, the effect on voicing boundaries of manipulating visual place of articulation closely parallels the effect of manipulating its auditory counterpart (Green & Kuhl, 1989). However, it is not presently known whether the common influence of auditory and visual place of articulation is limited to their effects on voicing boundaries or, instead, extends to the best-exemplar ranges within voiceless consonant categories.

We addressed this question in the present set of experiments. To do so, we built on two findings from the literature: Green and Kuhl's (1989) finding that voicing boundaries shift along a VOT continuum with a change in perceived place of articulation, owing to information in the visual signal, and Volaitis and Miller's (1992) finding that auditory place of articulation affects the locations of the best-exemplar ranges of voiceless consonants along a VOT continuum. We investigated whether a change in perceived place of articulation, attributable solely to a change in the visual signal, would cause a shift in the best-exemplar range for voiceless consonants. Specifically, we presented the stimuli from an auditory /bi/–/pi/–*/pi/ continuum with a visual /pi/ (perceived as members of a /bi/– /pi/–*/pi/ continuum) in one condition, and the same auditory stimuli with a visual /ti/ (perceived as members of a /di/–/ti/–*/ti/ continuum) in another condition and asked subjects to rate the stimuli as exemplars of /p/ in the first condition and as examplars of /t/ in the second. By comparing the best-exemplar ranges for /p/ and /t/ in the two conditions, we tested whether the effect of visual place of articulation, reported for voicing boundaries by Green and Kuhl, would extend to the best-exemplar range of the voiceless category, similar to the effect of auditory place of articulation. If internal category structure is sensitive to visual, as well as to auditory, information for place of articulation, the best exemplars for perceived /t/ (when the visual signal is /ti/) should fall over a range of longer VOTs than do the best exemplars of /p/ (when the visual signal is /pi/). That is, both the lower and the upper limits of the best-exemplar range for /t/ should occur at longer VOTs than do those for /p/. In contrast, if internal category structure is not sensitive to visual contextual information, the outcome might be similar to that of Allen and Miller's (2001) lexical manipulation, with only the portion of the best-exemplar range closest to the voicing boundary being affected, without a shift in the entire range. That is, there would be a shift in the lower limit of the best-exemplar range, but not in the upper limit.

We report the results of two experiments. Experiment 1 was a preliminary experiment designed to demonstrate a shift in the best-exemplar range for /p/ versus /t/ when the difference in place of articulation is auditory. Experiment 2 addressed our main question regarding the effect of visual place of articulation. It tested whether there is a similar shift in the best-exemplar range for /p/ versus /t/ when the change in perceived place of articulation occurs as a consequence of a change in the visual signal, in the absence of a change in the auditory place of articulation.

EXPERIMENT 1

The primary goal of Experiment 1 was to confirm that the differences in internal category structure between /p/ and /k/ observed by Volaitis and Miller (1992) are also found between /p/ and /t/. Thus, we conducted an auditory-only experiment involving presentation of stimuli from /bi/–/pi/–*/pi/ and /di/–/ti/–*/ti/ continua, in which subjects rated the consonants in the former set of stimuli as exemplars of /p/ and those in the latter set as exemplars of /t/. We expected that both the lower and the upper limits of the best-exemplar range for /t/ (like the range for /k/) would occur at longer VOTs than would the corresponding limits of the range for /p/, consistent with findings that the /d/–/t/ voicing boundary (like the /g/–/k/ boundary) occurs at a longer VOT than the /b/–/p/ boundary (Lisker & Abramson, 1970).

A secondary purpose of Experiment 1 was to test whether this goodness-rating paradigm could be successfully applied to audiovisual stimuli. We wanted to confirm that subjects' processing of fine-grained aspects of the auditory signal would not be disrupted by concurrent presentation of a visual stimulus. To test this, we conducted a second goodness experiment in which we presented the auditory stimuli from the /bi/–/pi/–*/pi/ and /di/–/ti/–*/ti/ continua with visual stimuli that matched the auditory stimuli in place of articulation; in other words, the bilabial auditory stimuli were presented with a bilabial visual stimulus (/pi/), and the alveolar auditory stimuli were presented with an alveolar visual stimulus (/ti/).[1]

Thus, Experiment 1 consisted of an auditory-only experiment (Experiment 1A) involving comparison of ratings to stimuli from auditory /bi/–/pi/–*/pi/ and /di/–/ti/–*/ti/ continua and an audiovisual experiment (Experiment 1B) involving comparison of ratings to the same auditory stimuli paired with congruent visual stimuli. We required two findings in order to proceed to our main experiment (Experiment 2). The first requirement was that, in the auditory-only experiment, both the lower and the upper limits of the best-exemplar range for /ti/ (as measured in the goodness-rating task) occurred at longer VOTs than did those of the range for /pi/. The second requirement was that the addition of a visual signal did not interfere with the subjects' ability to perform the goodness task, ev-

idenced by the presence of orderly goodness functions in the audiovisual experiment with shifts in both the lower and the upper limits of the best-exemplar range for /t/ relative to that for /p/, as in the auditory-only experiment.

Method

Subjects

Fourteen members of the Northeastern University community participated in the auditory-only experiment (1A), and another 14 participated in the audiovisual experiment (1B). 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

In Experiment 1A, the stimuli consisted of two auditory series of syllables varying in VOT; one ranged from /bi/ to /pi/ to a breathy exaggerated version of /pi/ (*/pi/), and the other ranged from /di/ to /ti/ to a breathy exaggerated version of /ti/ (*/ti/). In Experiment 1B, the stimuli from the same auditory /bi/–/pi/–*/pi/ and /di/–/ti/–*/ti/ series were paired with a visual token of a mouth producing /pi/ or /ti/, respectively.

Auditory stimuli. The auditory continua were created using the following steps.

Step 1. One clearly articulated token of /bi/ and one clearly articulated token of /di/, spoken by a female native speaker of American English, served as base stimuli for the synthesized auditory continua. The utterances were recorded via a microphone (AKG C460B) onto digital audio tape (TASCAM DA-P1 DAT recorder) in a sound-treated room and were transferred to a Pentium PC at a sampling rate of 20 kHz, using the CSL system (Kay Elemetrics Corp.). The durations of the /bi/ and the /di/ tokens were 522 and 536 msec, respectively. With the ASL program (Kay Elemetrics Corp.), a pitch-synchronous LPC analysis was performed on the waveforms of the /bi/ and /di/ tokens, using the autocorrelation method with a filter order of 16. The procedure extracts parameters for peak amplitude, fundamental frequency ($F0$), and formant frequencies and bandwidths, along with a residual excitation, for each frame. For these stimuli, each frame captured a single pitch period, with the exception of the first frame, which was set to include the initial burst and voiceless aspiration. The durations of the first frame of the /bi/ and the /di/ tokens were 6 and 17 msec, respectively.

Step 2. Two series of stimuli (one /bi/–/pi/–*/pi/ and one /di/–/ti/–*/ti/) were created by systematically changing parameters on a frame-by-frame basis and synthesizing new stimuli using the modified parameters. The first token in each series was created from synthesis, using the originally extracted parameters. The next token was created by changing the excitation parameter from the residual to a noise source and the $F0$ parameter to 0 (signifying voicelessness) in the first voiced frame and by scaling the peak amplitude of the frame by .15. (The parameters of the first frame in each series, which captured the initial burst and aspiration, were not modified.) The rest of the stimuli in each series were created by repeating this procedure for progressively increasing numbers of voiced frames. This was repeated until 42 continuum steps were created. In the /bi/–/pi/–*/pi/ series, the steps ranged in VOT from 6 to 206 msec, whereas the steps in the /di/–/ti/–*/ti/ series ranged in VOT from 17 to 214 msec. The step size was approximately 5 msec in both series. This procedure created a series of syllables ranging perceptually from /bi/ to /pi/ to a breathy exaggerated /pi/ (*/pi/) and a series ranging perceptually from /di/ to /ti/ to a breathy exaggerated /ti/ (*/ti/).

Step 3. Because the stimuli in the bilabial and the alveolar continua had different overall durations, we truncated each stimulus at 300 msec, using the waveform editor in CSL.[2] A descending cosine ramp was applied over the final 30 msec of each syllable in order to simulate a realistic amplitude contour.

Step 4. So that the release burst in each auditory token would be temporally aligned with the release in each of the visual tokens in Experiment 1B (described below), 433 msec of silence was added to each syllable prior to the consonant burst. Finally, the sound files for each syllable were converted to a sampling rate of 22.05 kHz (for presentation on a Macintosh G3).

Step 5. A subset of 20 of the 42 stimuli in each continuum was selected for use in the experiments. Every other step in the continua was selected, resulting in step sizes of approximately 10 msec, and the steps were chosen so that the VOTs in the /bi/–/pi/–*/pi/ and /di/–/ti/–*/ti/ continua would be as closely matched as possible. The VOTs of the selected stimuli in the /bi/–/pi/–*/pi/ series (in milliseconds) were 20, 30, 40, 50, 60, 69, 79, 89, 98, 108, 118, 127, 137, 147, 156, 166, 176, 186, 196, and 206. The VOTs of the selected stimuli in the /di/–/ti/–*/ti/ series (in milliseconds) were 21, 31, 41, 51, 60, 70, 79, 89, 99, 108, 118, 127, 137, 146, 156, 166, 175, 185, 195, and 204.

Visual stimuli. The visual /pi/ and /ti/ stimuli used in the audiovisual experiment (1B) were created using the following steps.

Step 1. The same speaker who was recorded for the auditory tokens produced multiple tokens of the utterances /əpi/ and /əti/. The initial schwa was included so that both the movement into the consonant closure and the consonant release would be clearly visible. The speaker's speech was recorded on videotape, using a Panasonic AG-188 VHS videocamera. Her mouth was fully illuminated. 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, using Adobe Premiere, at 30 frames per second at a resolution of 320 × 240 pixels. One clearly articulated token of /əpi/ and one clearly articulated token of /əti/ were selected.

Step 2. The visual tokens of /əpi/ and /əti/ were edited in Adobe Premiere in order to achieve two goals. The first goal was for the two visual tokens to be temporally aligned at their consonant release and to be matched in overall duration. The second goal was for each token to appear as though the speaker initiated the utterance from a neutral open-mouth position prior to the consonant's onset, instead of appearing to produce a preceding vowel (/ə/). Video frames were removed from the beginning of each token so that the point of consonant release (determined by a frame-by-frame analysis of the visual signal and by visual inspection of the acoustic waveform for the utterance, which was time-locked to the visual signal) occurred in the 9th frame. With this editing procedure, the resulting onset of each visual token occurred in the vicinity of the maximal jaw opening of the initial /ə/. Next, five repetitions of the initial frame were included at the stimulus onset, so that the utterance appeared to start from a static resting position. Consequently, the consonant release occurred in the 14th frame (433 msec after stimulus onset). Because the initial vowel /ə/ was no longer identifiable as a vowel after these editing procedures, we will refer to the video tokens as /pi/ and /ti/ (rather than /əpi/ and /əti/). Finally, the stimulus offset in each token was truncated so that each stimulus was 39 frames (1,300 msec) long. Each visual token was saved without an auditory channel.

Audiovisual alignment. In Experiment 1B, cross-modal alignment of the auditory and visual stimuli was achieved on line by presenting a sound file and a video file simultaneously (see the Procedure section). Because of the manner in which the stimuli were constructed, simultaneous presentation resulted in the visual stimulus's onset occurring 433 msec prior to the consonant onset in the auditory stimulus, so that the visible consonant release coincided with the acoustic consonant onset. The visual syllables had considerably longer durations (867 msec) than did the auditory syllables (300 msec), meaning that the acoustic offset occurred while the visible mouth was still open.

Preliminary study. As a preliminary step to our goodness experiments, we ran two forced-choice identification experiments in order to confirm that our bilabial and alveolar auditory series would

produce the standard voicing boundary shift with a change in place of articulation (e.g., Lisker & Abramson, 1970) and also that this shift would occur when the auditory stimuli were presented simultaneously with a congruent visual stimulus. In the auditory-only experiment, 10 subjects (none of whom participated in the goodness experiments) were presented with the auditory stimuli from the /bi/–/pi/–*/pi/ continuum in one block and the auditory stimuli from the /di/–/ti/–*/ti/ continuum in another block, and were asked to identify the initial consonant in each token as /b/, /p/, /d/, or /t/. In the audiovisual experiment run with 10 new subjects (none of whom participated in the goodness experiments), the same stimuli and procedure were used, except that the auditory stimuli from the /bi/–/pi/–*/pi/ continuum were presented with the visual token of /pi/, and the auditory stimuli from the /di/–/ti/–*/ti/ continuum were presented with the visual token of /ti/. The order of the two blocks was counterbalanced across subjects in both experiments. The voiced–voiceless category boundary was determined for each subject for each series by fitting a normal ogive to the identification data (excluding /d/ and /t/ responses to the /bi/–/pi/–*/pi/ stimuli and /b/ and /p/ responses to the /di/–/ti/–*/ti/ stimuli, which represented fewer than 2% of the responses) and calculating the mean of the ogive function, corresponding to the VOT (in milliseconds) at which voiced and voiceless responses were equally probable. In the auditory-only experiment, the mean /d/–/t/ boundary (66-msec VOT) occurred at a longer VOT than did the mean /b/–/p/ boundary (43-msec VOT). The audiovisual experiment produced nearly identical results, with the mean /d/–/t/ boundary (67-msec VOT) occurring at a longer VOT than did the mean /b/–/p/ boundary (44-msec VOT). The difference was highly significant in both experiments [auditory-only experiment, $t(9) = 9.78$, $p < .0001$; audiovisual experiment, $t(9) = 8.20$, $p < .0001$].

Procedure

In both Experiments 1A and 1B, the stimuli were presented via a Macintosh G3 located in a sound-treated booth. Stimulus presentation was controlled by PsyScope (Cohen, MacWhinney, Flatt, & Provost, 1993). The auditory syllables were presented binaurally through Sony MDR-V6 headphones at a comfortable listening level that remained constant throughout the experiment. In the audiovisual experiment (1B), the visual stimuli were presented on a 17-in. monitor. The video image filled about three quarters of the screen and was surrounded by a black screen. In both experiments, the subjects were seated in front of the computer monitor at a normal viewing distance (approximately 18 in.). The experimenter observed the experiments through a window in the booth to confirm that the subjects attended to the visual presentations.

Both Experiments 1A and 1B were run in two separate blocks, each of which involved presentation of either the bilabial or the alveolar stimuli. The order of these blocks was counterbalanced across subjects. Each block began with a familiarization phase, in which the stimuli from the auditory continuum were presented in order of increasing VOT; the sequence of sounds was played twice. The subjects were told that the sounds would progress from /b/ to /p/ to a breathy exaggerated /p/ (or /d/ to /t/ to a breathy exaggerated /t/), and they then listened to the sequence of syllables without responding. In the audiovisual experiment (1B), the auditory syllables were presented with a visual token of the syllable with a matching place of articulation; thus, the auditory syllables from the /bi/–/pi/–*/pi/ series were presented with a visual /pi/, and the auditory syllables from the /di/–/ti/–*/ti/ series were presented with a visual /ti/.

In the test phase of the experiments, the subjects were told to rate each auditory syllable as an exemplar of a particular consonant (/p/ for the bilabial series and /t/ for the alveolar series), using a scale from 1 (corresponding to a *poor exemplar*) to 7 (corresponding to a *very good exemplar*). The subjects were encouraged to use the entire scale in making their judgments. In the audiovisual experiment (1B), the subjects were instructed to watch the video presentation on the screen but, nonetheless, to base their judgments on the sound.

Each trial in the experiments began with a warning tone, and after a 500-msec pause, an auditory stimulus was played (with 433 msec of silence at its onset). In the audiovisual experiment (1B), a visual stimulus was presented simultaneously with the auditory stimulus. After the offset of the stimulus, seven boxes, labeled 1 through 7, appeared on the screen underneath the query "How good of a 'P'?" (or "How good of a 'T'?"), which disappeared after the subject had responded by clicking in one of the boxes. The next trial began after the subject moved the cursor to a visually designated location at the bottom of the screen. The screen-based responding and trial initiation procedures were intended to maintain the subjects' visual attention on the computer monitor throughout the experiments.

Each test block was preceded by seven practice trials, in which stimuli that were roughly evenly spaced across the VOT continuum were presented in a random sequence. Each test block involved 13 randomized sequences of the 20 auditory tokens from one of the auditory series (either /bi/–/pi/–*/pi/ or /di/–/ti/–*/ti/). The subjects received breaks approximately halfway through each block and between the two blocks. The first 20 trials of each block were not included in the data analysis.

Data Analysis

Mean goodness ratings across the 12 repetitions of each auditory stimulus were computed separately for each subject, generating individual goodness-rating functions for the /bi/–/pi/–*/pi/ and the /di/–/ti/–*/ti/ series. These individual functions were then smoothed by computing an average of the rating for each continuum step with the ratings for the two adjacent continuum steps. The endpoints of the continua were not affected by the smoothing procedure. The smoothed data provided the basis for all of the analyses reported in this paper, although the figures present the unsmoothed data. In general, the shape of the resulting functions was typical of experiments using VOT continua with the goodness-rating paradigm: Ratings were low for short VOTs, increased with increasing VOT until they reached a maximal level, and then decreased as VOT became longer.

For each subject, we determined the location of the *best-exemplar range* for the goodness function for the /bi/–/pi/–*/pi/ continuum and for the /di/–/ti/–*/ti/ continuum. The best-exemplar range of each function was quantified as the range of VOTs for which the ratings were at least 90% of the peak rating for that function (see Allen & Miller, 2001; Miller & Volaitis, 1989; Volaitis & Miller, 1992). The lower and upper *limits* of the best-exemplar ranges were established in the following manner, separately for each function. First, the rating corresponding to 90% of the peak rating (the highest mean rating of any stimulus along the continuum) was determined. Next, the two points at which the goodness function (one at a shorter VOT than the stimulus with the peak rating and one at a longer VOT) crossed this 90% value were determined. This was accomplished by linear interpolation between the continuum steps with ratings that straddled the 90% value.

Subject Criteria

In goodness-rating experiments such as the ones reported here, a few subjects typically experience difficulty with the task and fail to produce orderly goodness functions (Allen & Miller, 2001; Miller et al., 1997; Wayland et al., 1994). However, our analysis presumes that the goodness functions are orderly; if they are not, the lower- and upper-limit measures are computationally not valid. Therefore, we set criteria for excluding any subjects with atypical goodness functions, resulting in the elimination of 3 subjects (1 in the auditory-only experiment and 2 in the audiovisual experiment) from our analyses. Consequently, the data reported are based on the results of 13 subjects in the auditory-only experiment (1A) and 12 subjects in the audiovisual experiment (1B).

Our criteria were as follows. First, we eliminated from all analyses any subject whose goodness functions (for either /p/ or /t/) did not drop to 90% of the peak value at either a long or a short VOT, because the 90% crossover point was required for our quantification

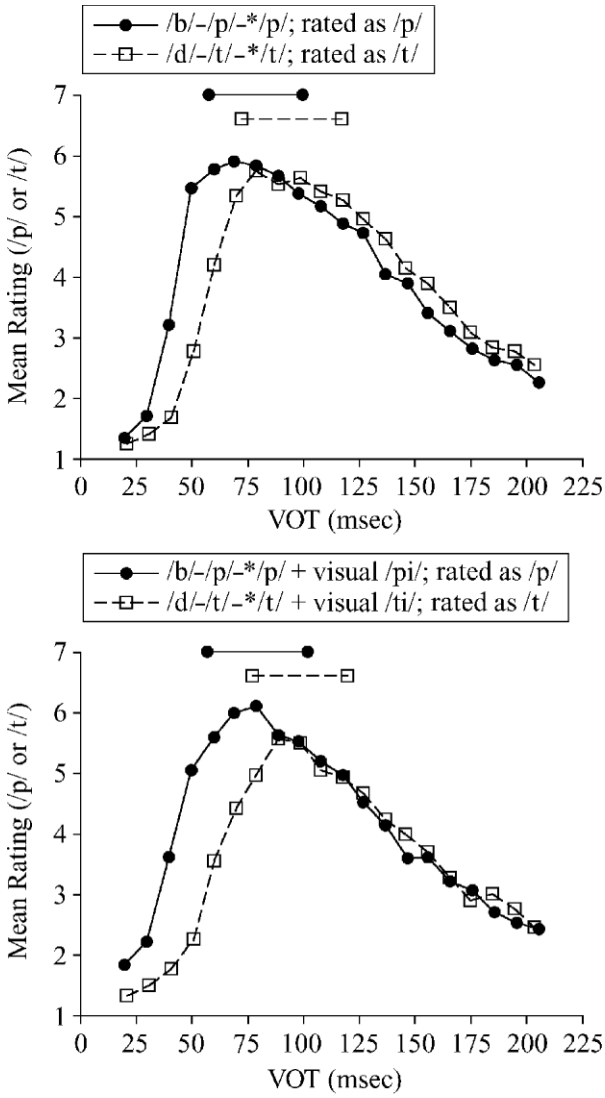

Figure 1. Group goodness ratings as a function of voice onset time (VOT) for the auditory /bi/–/pi/–*/pi/ continuum, rated as /p/, and the /di/–/ti/–*/ti/ continuum, rated as /t/, in Experiment 1. The top panel presents the mean ratings for the auditory-only condition (Experiment 1A), and the bottom panel presents the mean ratings for the audiovisual condition (Experiment 1B). The solid and dashed horizontal lines at the top of each figure represent the best-exemplar ranges for /p/ and /t/, respectively.

of the best-exemplar range. Second, because the stimuli with very short VOTs should have been perceived as voiced consonants (either /b/ or /d/, depending on the continuum) and, therefore, should have received low ratings as an exemplar of a voiceless consonant (either /p/ or /t/), we also eliminated the data of any subject whose ratings did not fall below 75% of the peak rating at a short VOT for either /p/ or /t/. Finally, we found that some subjects produced broad best-exemplar ranges without a well-defined peak and without a clear-cut dropoff in ratings for long VOTs. Because we did not consider the upper limits to be meaningful for such functions, we eliminated any subjects for whom the width of the best-exemplar range (for either /p/ or /t/) was greater than three standard deviations above the mean width.

Results and Discussion

Auditory-Only Experiment (Experiment 1A)

The group (unsmoothed) goodness functions for /p/ and /t/ in the auditory-only experiment are shown in the top panel of Figure 1. As the figure demonstrates, both functions progress in an orderly fashion with increasing VOT: Ratings are very low for the stimuli with the shortest VOTs, increase rather sharply with increasing VOT until they reach a maximal level, and then decrease gradually as VOT becomes longer. The figure also demonstrates that although ratings were overall somewhat lower for /ti/ than for /pi/, the entire function for /ti/ appears to be shifted to longer VOTs in comparison with the /pi/ function, with a concomitant shift in the range of VOTs that received the highest ratings for /t/ relative to the comparable range for /p/.[3]

Quantitative support for this difference in the location of the best-exemplar ranges for /p/ and /t/ is provided by our measures of the lower and upper limits of these ranges (see the Data Analysis section above). The best-exemplar ranges are represented as horizontal bars in Figure 1. Both the lower and the upper limits for /t/ (lower, $M = 72.5$ msec, $SE = 3.5$; upper, $M = 117.4$ msec, $SE = 4.1$) occurred at longer VOTs than did those for /p/ (lower, $M = 57.7$ msec, $SE = 3.0$; upper, $M = 99.9$ msec, $SE = 5.2$). A 2×2 repeated measures analysis of variance (ANOVA) with rated consonant (/p/ vs. /t/) and limit (lower vs. upper) as factors, revealed that these differences were significant: There was a significant main effect of rated consonant [$F(1,12) = 82.90$, $p < .0001$], as well as a significant main effect of limit [$F(1,12) = 551.28$, $p < .0001$], but no interaction between the two factors [$F(1,12) < 1$, n.s.]. The lack of an interaction indicates that /p/ and /t/ differed at both the lower and the upper limits of the best-exemplar range (lower limit, 14.8-msec shift; upper limit, 17.5-msec shift).

Audiovisual Experiment (Experiment 1B)

The group (unsmoothed) goodness functions for /p/ and /t/ in the audiovisual experiment are shown in the bottom panel of Figure 1. Overall, the pattern of results is very similar to that found in the auditory-only experiment. First, the subjects produced orderly goodness functions with a progression from low to high to low ratings as VOT increased. (Note, also, that only 1 more subject was eliminated for having atypical functions in the audiovisual condition than in the auditory-only condition, suggesting that the difficulty of the goodness task did not differ substantially across the two conditions.) Second, although again the ratings were somewhat lower for /t/ than for /p/, the highest ratings for /t/ occurred over a range of longer VOTs than did the highest ratings for /p/; this is demonstrated in the figure by the horizontal bars representing the best-exemplar ranges for /p/ and /t/.

As in the auditory-only experiment, both the lower and the upper limits for /t/ (lower, $M = 77.2$ msec, $SE = 3.6$; upper, $M = 119.9$ msec, $SE = 5.3$) occurred at longer VOTs than did those for /p/ (lower, $M = 57.0$ msec, $SE = 4.9$; upper, $M = 102.0$ msec, $SE = 5.8$). The results of sta-

tistical analyses provided an outcome that closely paralleled that of the auditory-only experiment: A 2 × 2 repeated measures ANOVA, with rated consonant (/p/ vs. /t/) and limit (lower vs. upper) as factors, revealed significant main effects of both rated consonant [$F(1,11) = 58.01, p < .0001$] and limit [$F(1,11) = 560.98, p < .0001$], but no interaction between the two factors [$F(1,11) < 1$, n.s.]. Thus, /p/ and /t/ differed at both the lower and the upper limits of their respective best-exemplar ranges (lower limit, 20.2-msec shift; upper limit, 17.9-msec shift).

Finally, we compared the results of the auditory-only and the audiovisual experiments in a 2 × 2 × 2 ANOVA, with experiment (auditory only vs. audiovisual), rated consonant (/p/ vs. /t/), and limit (lower vs. upper) as factors. The main effect of experiment and all higher order interactions involving experiment failed to approach significance ($F < 1$ in all cases). Thus, the lower and upper limits for /p/ and /t/ were not affected by the presentation of a congruent visual token.

Summary

Experiment 1 was conducted to assess two prerequisite conditions for Experiment 2. The first condition was that, in the auditory-only experiment, /p/ and /t/ would differ in their best-exemplar ranges, following Volaitis and Miller's (1992) finding of a difference between the best-exemplar ranges for /p/ and /k/. The second condition was that the subjects could effectively perform the goodness task, which requires attention to fine-grained aspects of the auditory signal, while simultaneously attending to a visual signal. The results were clear in demonstrating that both conditions were met. First, in the auditory-only experiment, the entire best-exemplar range for /t/ spanned a region covering longer VOTs than did the corresponding range for /p/, evidenced by shifts at both the lower and the upper limits of the best-exemplar range. Second, we observed orderly goodness functions in the audiovisual experiment, with differences between the best-exemplar ranges for /p/ and /t/ that were comparable to those observed in the auditory-only experiment. Given these findings, we proceeded to Experiment 2.

EXPERIMENT 2

The purpose of Experiment 2 was to test whether a change in visual place of articulation from bilabial to alveolar, without an associated change in auditory place of articulation, produces a shift in the best-exemplar range for voiceless consonants that is similar to the shift produced by a change in auditory place of articulation. As in Experiment 1, we compared the lower and the upper limits of the best-exemplar ranges for /p/ and /t/, using auditory continua, but instead of using auditory /bi/–/pi/–*/pi/ and /di/–/ti/–*/ti/ continua in the two conditions, we used only an auditory /bi/–/pi/–*/pi/ continuum, presented with a congruent visual /pi/ in one condition and an incongruent visual /ti/ in the other condition. Whereas the congruent visual /pi/ should not alter the perceived place of articulation of the auditory stimuli, the incongruent visual /ti/ should cause the auditory stimuli to be perceived as members of a /di/–/ti/–*/ti/ continuum. Accordingly, the subjects were instructed to rate the stimuli in the congruent condition as exemplars of /p/ and the stimuli in the incongruent condition as exemplars of /t/.

It is important to note that the goodness task in the incongruent condition presumes that subjects experience a McGurk effect—that is, they perceive the stimuli as either /d/ or /t/. However, previous research has shown that subjects may exhibit the McGurk effect only on a proportion of trials when presented with a given set of audiovisually incongruent stimuli (Brancazio, 2003; Green & Norrix, 1997; MacDonald & McGurk, 1978; Massaro, 1998), and the overall incidence of the effect varies across individuals (Brancazio, Miller, & Paré, 1999; Carney, Clement, & Cienkowski, 1999). Our goodness task does not provide subjects with an opportunity to report whether they have perceived an alveolar or a bilabial consonant on each trial, although, as we describe in the Method section, our subjects readily accepted our description of the incongruent stimuli as belonging to an auditory series ranging from /di/ to /ti/ to a breathy exaggerated /ti/. Nonetheless, because we lacked a direct measure of whether the subjects perceived an alveolar consonant on each trial, we restricted our analysis of the goodness ratings to individuals who provided independent evidence of experiencing a strong McGurk effect with our stimuli. To do so, we conducted a phoneme identification pretest with all of our subjects, in which they identified the stimuli used in the congruent and incongruent conditions of our goodness-rating experiment as /b/, /p/, /d/, or /t/, and we established an inclusion criterion (described in the Method section) for our subjects based on how often they identified the incongruent stimuli as /d/ or /t/. Note that this identification pretest was essentially a replication of Green and Kuhl's (1989) study, with two differences. First, our control condition involved audiovisually congruent stimuli, whereas Green and Kuhl's involved only auditory stimuli; and second, our incongruent visual token was an alveolar (/t/) rather than a velar (/g/) consonant, although, for both, the typical McGurk percept is an alveolar consonant. Accordingly, in addition to providing a screening measure, the identification pretest served another purpose: It enabled us to compute voicing boundaries for the congruent and the incongruent conditions, providing a test of whether Green and Kuhl's findings (i.e., a voicing boundary shift to a longer VOT in the incongruent condition) would replicate with different stimuli.

The predictions for the goodness-rating experiment were straightforward. If the internal category structure of voiceless consonants is sensitive to visual information for place of articulation, the entire best-exemplar range for /t/ in the incongruent condition should occur over longer VOTs than does the best-exemplar range for /p/ in the congruent condition, with shifts in both the lower and the upper limits of the range. If, however, effects of visual place of ar-

ticulation on voicing are only a boundary phenomenon, the best-exemplar range for /t/ in the incongruent condition would differ from that for /p/ in the congruent condition, if at all, only in the portion closest to the voicing boundary. This would yield a shift at the lower, but not at the upper, limit of the range, as in Allen and Miller (2001).

Method

Subjects

Twenty different members of the Northeastern University community participated in the experiment. 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 consisted of the auditory tokens from the /bi/–/pi/–*/pi/ continuum and the visual /pi/ and /ti/ tokens used in Experiment 1.

Procedure

The experiment consisted of a phoneme identification pretest and a goodness-rating task. All the subjects performed both tasks in separate sessions conducted over 3 days, with the first session devoted to the identification pretest and the second and third sessions devoted to the goodness-rating task. Each session took approximately 1 h to complete. The experimental setting for all three sessions was the same as that used in Experiment 1.

On each trial in the phoneme identification task, a token from the /bi/–/pi/–*/pi/ continuum was presented simultaneously with a visual stimulus (either /pi/ or /ti/), preceded by a warning tone and a 500-msec pause. After the offset of the stimulus, the subjects responded by clicking on one of four boxes, labeled "B," "P," "D," and "T," that appeared on the screen. 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, regardless of what they saw.

The identification task consisted of two blocks. In the first block, the *incongruent* condition, the visual stimulus was always /ti/ (thus differing from the auditory stimuli in place of articulation) and included 21 randomized sequences of the 20 auditory tokens in the /bi/–/pi/–*/pi/ series paired with the incongruent video. The first 20 trials served as familiarization trials and were not included in the data analysis. In the second block, the *congruent* condition, the visual stimulus was always /pi/ (thus matching the auditory stimuli in place of articulation) and included 10 randomized sequences of the 20 auditory tokens with the congruent video. The incongruent block was always presented first, in order to provide us with a consistent measure of the magnitude of the McGurk effect for each subject, without a potential confound of prior experience with the stimuli. We ran twice as many trials with the incongruent video as with the congruent video for the purposes of computing a /d/–/t/ boundary for each subject; we anticipated that many subjects would respond "b" or "p" on a substantial number of trials, and we wanted to ensure that we would have a sufficient number of /d/ and /t/ responses for each subject. The subjects received a break halfway through the first block and a second break between the first and the second blocks.

The goodness-rating task was run in the same manner as in Experiment 1, with a familiarization phase, a set of seven practice trials, and a test phase. The two conditions (run on 2 separate days, with the order counterbalanced across subjects) involved presentation of the sounds from the /bi/–/pi/–*/pi/ continuum, presented in conjunction with a visual token. However, a different visual token was used in each: In the *congruent* condition, the visual token was the audiovisually congruent /pi/, and in the *incongruent* condition, the visual token was the audiovisually incongruent /ti/. The instructions used in both conditions of the goodness-rating experiment were closely matched to the instructions used in the respective conditions of Experiment 1B. In the congruent (/p/) condition, the subjects were informed prior to the familiarization phase that they would hear a series of sounds ranging from /bi/ to /pi/ to */pi/, presented in conjunction with a visible articulating mouth. After the familiarization trials, they were instructed to rate each stimulus as an exemplar of the consonant /p/, using a scale ranging from 1 (*poor*) to 7 (*very good*). In the incongruent (/t/) condition, the subjects were informed in the familiarization phase that they would hear a series of sounds ranging from /di/ to /ti/ to */ti/, presented in conjunction with a visible articulating mouth. They were then instructed to rate each stimulus as an exemplar of the consonant /t/, using the scale ranging from 1 to 7. In both sessions, the subjects made their response by clicking on numbered boxes on the screen, accompanied by a prompt of either "How good of a 'P'?" or "How good of a 'T'?" The subjects were not informed that the stimuli were the same as the ones used in the prior phoneme identification task. Furthermore, the instructions indicated to the subjects that the auditory signals in the incongruent condition were actually alveolar consonants. These instructions appeared to be effective: None of our subjects, including those with a relatively low incidence of the McGurk effect on the identification task, objected to our characterization of the incongruent stimuli as ranging from /di/ to /ti/ to a breathy exaggerated /ti/.

Each test block consisted of 21 randomized sequences of the 20 auditory tokens; the subjects received two breaks during each session. The first 20 trials of each block were not included in the data analysis.

Data Analysis

Phoneme identification pretest. The subjects' responses on the phoneme identification pretest were used for two purposes. One purpose of the pretest was to provide a screening measure to enable us to limit the analysis of the goodness results to subjects who experienced a strong McGurk effect. Accordingly, the overall magnitude of the McGurk effect was determined for each subject by computing the percentage of visually influenced, or McGurk, responses for all of the trials (collapsing across continuum steps) in the incongruent condition.

The other purpose of the pretest was to attempt to replicate Green and Kuhl's (1989) finding of a visually induced voicing boundary shift, by computing the voicing boundaries between /b/ and /p/ in the congruent condition (with visual /pi/) and between /d/ and /t/ in the incongruent condition (with visual /ti/). Voicing boundaries were computed in the following manner. First, responses that did not match the visual token in place of articulation were eliminated; that is, only /b/ and /p/ responses (eliminating /d/ and /t/ responses) in the congruent condition and only /d/ and /t/ responses (eliminating /b/ and /p/ responses) in the incongruent condition were used to compute voicing boundaries. Overall, fewer than 1% of the trials in the congruent condition and approximately 21% of the trials in the incongruent condition were eliminated. Next, the percentage of voiceless responses in each condition (i.e., the percentage of /b/ and /p/ responses that were /p/ in the congruent condition, ignoring /d/ and /t/ responses to these stimuli, and the percentage of /d/ and /t/ responses that were /t/ in the incongruent condition, ignoring /b/ and /p/ responses) were computed for each continuum step. Finally, 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 function, corresponding to the VOT (in milliseconds) at which voiced and voiceless responses were equally probable.

Goodness ratings. Goodness ratings were analyzed in the same manner as in Experiment 1. For each subject, two goodness functions, one for the congruent condition (with visual /pi/) and one for the incongruent condition (with visual /ti/), were generated by com-

puting the mean rating across the 20 repetitions of a given continuum step and then applying the smoothing algorithm. As in Experiment 1, the locations of the lower and upper limits of the best-exemplar range for each goodness function were determined (in milliseconds of VOT).

Subject Criteria

As in Experiment 1, we eliminated the data of any subjects who failed to produce orderly goodness functions. The criteria from Experiment 1 were again employed, resulting in the elimination of 3 subjects. In addition, we restricted our analysis to those subjects who produced a robust McGurk effect, determined on the basis of the results of the phoneme identification pretest. We eliminated any subject who gave McGurk responses (i.e., /d/ or /t/) on fewer than 50% of the trials in the incongruent condition of the pretest. Only 1 subject who was not eliminated by the earlier set of criteria failed to meet this criterion. Thus, a total of 4 subjects were eliminated from all of the analyses reported below. We note that of the 3 subjects who were eliminated because they produced atypical goodness functions, 2 would have been eliminated by the McGurk effect criterion as well. It is perhaps not surprising that the 2 subjects who did not experience a consistent McGurk effect exhibited difficulty with the goodness task, since the weak McGurk effect may have caused them to perceive the stimuli across the continuum as poor exemplars of /t/ in terms of place of articulation.

Results and Discussion

Phoneme Identification Pretest

The pretest data allowed us to determine whether we replicated Green and Kuhl's (1989) finding of a voicing boundary shift that was due to presentation of a visual stimulus that induced a change in perceived place of articulation. To do so, we compared the locations of the /b/–/p/ boundary in the audiovisually congruent condition (with visual /pi/) and the /d/–/t/ boundary in the incongruent condition (with visual /ti/). The mean /b/–/p/ boundary was 48.2-msec VOT, and the mean /d/–/t/ boundary was 67.3-msec VOT. A paired t test revealed that this difference was significant [$t(15) = 10.04, p < .0001$]. This boundary shift is clearly shown in Figure 2, which presents the mean percentages of voiceless responses (/p/ and /t/ for the congruent and the incongruent conditions, respectively) for each step of the continuum. Thus, we found that when auditory bilabial stimuli are presented with an incongruent video and are perceived as alveolar, the resulting voicing boundary occurs at a longer VOT relative to a baseline condition (with a congruent video). These results are consistent with those of Green and Kuhl, despite the methodological differences between the two studies, noted earlier.

Goodness Ratings

The results of the goodness-rating experiment allowed us to test whether the effect of visual information extends throughout the best-exemplar range. Figure 3 presents the group (unsmoothed) goodness functions for /p/ in the congruent condition and /t/ in the incongruent condition. The figure demonstrates that the subjects produced goodness functions in both conditions with the expected overall shape, with ratings systematically increasing and then decreasing as VOT increased. The figure also demonstrates that although ratings were slightly lower for /t/ than for /p/, the entire function for /t/, including its best-exemplar range (represented as a horizontal bar in the figure), was shifted toward longer VOT values relative to the function for /p/ and its best-exemplar range (also represented as a horizontal bar in the figure).

As in Experiment 1, we tested the extent of this shift by comparing the mean lower and upper limits of the best-exemplar range of each function. Both the lower and the upper limits of the best-exemplar range occurred at longer VOTs for /t/ (lower, $M = 76.9$ msec, $SE = 3.5$; upper, $M = 128.8$ msec, $SE = 5.5$) than for /p/ (lower, $M = 65.2$ msec, $SE = 2.7$; upper, $M = 110.2$ msec, $SE = 4.3$). In a 2×2 repeated measures ANOVA, with rated consonant (/p/, with a visual /pi/, vs. /t/, with a visual /ti/) and limit (lower vs. upper) as factors, we found significant main effects of both rated consonant [$F(1,15) = 29.3, p < .0001$] and limit [$F(1,15) = 237.1, p < .0001$], but no interaction between the two factors [$F(1,15) = 2.7, p > .10$]. The shift at the lower limit (11.7 msec) was somewhat smaller than the shift at the upper limit (18.6 msec), although both shifts were highly robust ($p < .005$ in each case). As a result of this difference, the best-exemplar range was slightly wider for /t/ than for /p/ (51.9 vs. 45.0 msec, observable as a slightly less peaked function for /t/ in Figure 3); however, this difference was not significant, as is indicated by the lack of a significant interaction between rated consonant and limit.

Summary

The results of Experiment 2 are clear in demonstrating that when stimuli from an auditory bilabial continuum are perceived as alveolar due to an incongruent visual token,

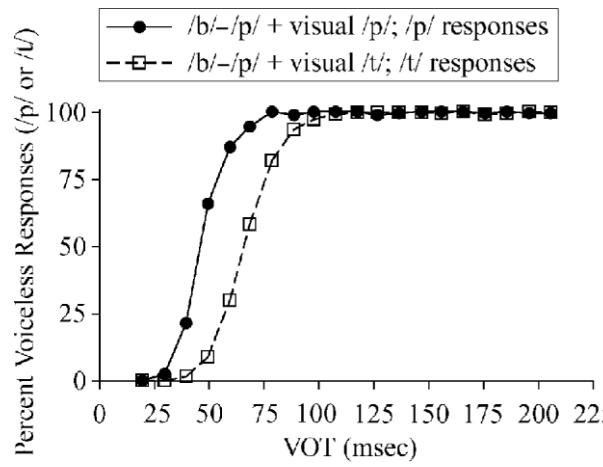

Figure 2. Group mean percentages of voiceless responses as a function of voice onset time (VOT) for the auditory /bi/–/pi/–*/pi/ continuum presented with a visual /pi/ in the congruent condition and presented with a visual /ti/ in the incongruent condition, in the phoneme identification pretest of Experiment 2. In the congruent condition, only /b/ and /p/ responses were considered, and the solid line presents the percentage of those responses that were /p/. In the incongruent condition, only /d/ and /t/ responses were considered, and the dashed line presents the percentage of those responses that were /t/.

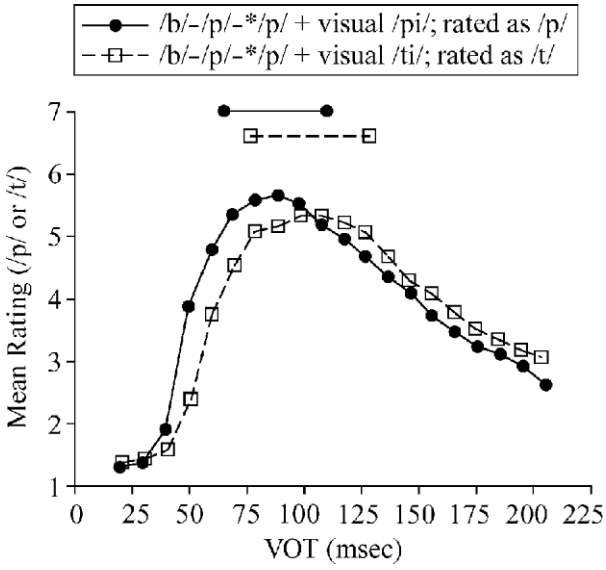

Figure 3. Group goodness ratings as a function of voice onset time (VOT) for the auditory /bi/–/pi/–*/pi/ **continuum presented with a visual** /pi/ **and rated as** /p/ **in the congruent condition and presented with a visual** /ti/ **and rated as** /t/ **in the incongruent condition, in Experiment 2. The solid and dashed horizontal lines at the top of the figure represent the best-exemplar ranges for** /p/ **and** /t/**, respectively.**

not only does the voicing boundary shift to a longer VOT, but also the best-exemplar range for the voiceless consonant shifts as well. Importantly, this effect entails shifts in both the lower and the upper limits of the best-exemplar range, rather than a shift only in the portion of the range closest to the voicing boundary. Thus, effects of visual place of articulation on voicing perception extend throughout the best-exemplar range for the voiceless category. In this regard, the effects of a change in visual place of articulation from bilabial to alveolar are comparable to the effects of a corresponding change in auditory place of articulation, as reported in Experiment 1.

GENERAL DISCUSSION

The purpose of the present research was to examine the effects of visual place of articulation on internal category structure for voiceless consonants. Specifically, we tested whether effects of visual place of articulation (previously reported for voicing boundaries by Green & Kuhl, 1989) would extend throughout the best-exemplar region, similar to the effects of auditory place of articulation (Volaitis & Miller, 1992) or, instead, would be limited to the voicing boundary region. The results of Experiment 2 clearly demonstrated that a change in the visual stimulus from /p/ to /t/, creating a change in perceived place of articulation from bilabial to alveolar, resulted in a systematic shift in the entire best-exemplar range for the voiceless consonant to longer VOTs. Moreover, this shift and the shift resulting from a change in auditory place of articulation from bilabial to alveolar (Experiment 1) were highly similar. Thus, the effect of visual place of articulation on voicing parallels the effect of auditory place of articulation.

This outcome is consistent with several findings demonstrating parallel effects of changes in auditory and visual context on phonetic boundaries. For example, Green and Miller (1985) found a shift in voicing boundaries due to a change in visual speaking rate (by presenting stimuli from an auditory /bi/–/pi/ continuum with visual tokens of a person saying /pi/ very quickly or very slowly), similar to the effect of a change in auditory speaking rate (Summerfield, 1981). Fowler, Brown, and Mann (2000) have also demonstrated that the phenomenon of compensation for coarticulation, illustrated by a shift in a /d/–/g/ boundary in the presence of a preceding /l/ versus /r/ (Mann, 1980), occurs with preceding visual /l/ and /r/ segments. In a similar vein, Green and Norrix (2001) found a boundary shift along an /l/–/r/ continuum that arose as a consequence of a preceding /b/ in either the auditory signal (presented dichotically) or the visual signal. Together, these findings suggest that when the visual signal provides information that is relevant for phonetic perception (including information about speaking rate, place of articulation, or coarticulatory consequences of neighboring segments), the speech perception system generally uses that information in a manner similar to that with which it uses comparable information in the auditory signal (although see Roberts & Summerfield, 1981; Saldaña & Rosenblum, 1994). The major contribution of the present findings in this regard is the demonstration that the common use of visual and auditory information is not limited to those processes that determine phonetic category boundaries but also extends to processes that determine the internal structure of phonetic categories.

The finding that visual information affects the internal structure of phonetic categories bears on the issue of which contextual factors influence internal category structure and which do not. Previous findings have suggested a dissociation between acoustic-phonetic factors, such as place of articulation and speaking rate, and higher order factors, such as lexical status, with only acoustic-phonetic contextual factors showing comprehensive effects on internal category structure (see Allen & Miller, 2001). To account for this dissociation, Allen and Miller proposed that a contextual factor affects the internal structure of a phonetic category along a particular acoustic dimension only if that factor also affects that dimension in the production of speech. In other words, according to this *production-based* account, contextual effects on best-exemplar ranges reflect the speech perception system's ability to track variation in speech production and to adjust its phonetic categories accordingly. This account provides an explanation for why internal category structure for voiceless consonants is sensitive to the contextual factors of place of articulation and speaking rate, which have systematic effects on VOT in speech production (Lisker & Abramson, 1964; Volaitis & Miller, 1992), but is not sensitive to lexical status, which does not systematically affect VOT (e.g., the VOTs for /p/ in *peace* and *peef* do not differ; Allen & Miller, 2001).

The present findings have important implications for this account because they necessitate a clarification of the distinction between acoustic-phonetic and higher order factors. Specifically, they demonstrate that the relevant properties that affect internal category structure need not be *acoustic*-phonetic per se; that is, they need not have a specifically auditory basis. Thus, for a production-based account to be viable, the perceptual system that tracks variation in speech production must use information in the visual, as well as in the auditory, signal. Because visual place of articulation is directly linked to the production of speech, the present findings are compatible with such a view. It is important to note that on this view, internal category structure will not be sensitive to any visual contextual factor, but only to those visual factors that specifically relate to speech production. Thus, future research in which the effects of other visual factors on internal category structure are examined can provide additional tests of the production-based account from a multimodal perspective.

REFERENCES

Allen, J. S., & Miller, J. L. (2001). Contextual influences on the internal structure of phonetic categories: A distinction between lexical status and speaking rate. *Perception & Psychophysics*, **63**, 798-810.

Brancazio, L. (2003). *Lexical influences in audiovisual speech perception*. Manuscript submitted for publication.

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

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

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.

Fowler, C. A., Brown, J. M., & Mann, V. A. (2000). Contrast effects do not underlie effects of preceding liquids on stop-consonant identification by humans. *Journal of Experimental Psychology: Human Perception & Performance*, **26**, 877-888.

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. (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., & 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.

Green, K. P., & Norrix, L. W. (2001). Perception of /r/ and /l/ in a stop cluster: Evidence of cross-modal context effects. *Journal of Experimental Psychology: Human Perception & Performance*, **27**, 166-177.

Hodgson, P., & Miller, J. L. (1996). Internal structure of phonetic categories: Evidence for within-category trading relations. *Journal of the Acoustical Society of America*, **100**, 565-576.

Kuhl, P. K. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics*, **50**, 93-107.

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.

Mann, V. A. (1980). Influence of preceding liquid on stop-consonant perception. *Perception & Psychophysics*, **28**, 407-412.

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

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

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

Miller, J. L. (1994). On the internal structure of phonetic categories: A progress report. *Cognition*, **50**, 271-285.

Miller, J. L., & Dexter, E. R. (1988). Effects of speaking rate and lexical status on phonetic perception. *Journal of Experimental Psychology: Human Perception & Performance*, **14**, 369-378.

Miller, J. L., O'Rourke, T. B., & Volaitis, L. E. (1997). Internal structure of phonetic categories: Effects of speaking rate. *Phonetica*, **54**, 121-137.

Miller, J. L., & Volaitis, L. E. (1989). Effect of speaking rate on the perceptual structure of a phonetic category. *Perception & Psychophysics*, **46**, 505-512.

Pitt, M. A. (1995). The locus of the lexical shift in phoneme identification. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **21**, 1-16.

Repp, B. H., & Liberman, A. M. (1987). Phonetic category boundaries are flexible. In S. Harnad (Ed.), *Categorical perception: The groundwork of cognition* (pp. 89-112). New York: Cambridge University Press.

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

Saldaña, H. M., & Rosenblum, L. D. (1994). Selective adaptation in speech perception using a compelling audiovisual adaptor. *Journal of the Acoustical Society of America*, **95**, 3658-3661.

Samuel, A. G. (1982). Phonetic prototypes. *Perception & Psychophysics*, **31**, 307-314.

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

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.

Wayland, S. C., Miller, J. L., & Volaitis, L. E. (1994). The influence of sentential speaking rate on the internal structure of phonetic categories. *Journal of the Acoustical Society of America*, **95**, 2694-2701.

NOTES

1. Because information about voicing is not available in the visual signal, a visually presented voiceless consonant (e.g., /pi/) will appear to be congruent with both voiced and voiceless auditory consonants that match it in place of articulation (e.g., /bi/ and /pi/).

2. We elected to shorten the stimuli by a considerable degree (e.g., 300 msec from original durations of over 500 msec) because previous work has indicated that short stimuli generate goodness functions with more sharply defined best-exemplar ranges than those of longer stimuli (e.g., Allen & Miller, 2001; Miller & Volaitis, 1989; Volaitis & Miller, 1992), which facilitates statistical comparison of best-exemplar ranges in different conditions.

3. Previous studies using the goodness-rating paradigm have also found disparities in the height of functions across conditions (Allen & Miller, 2001; Hodgson & Miller, 1996; Volaitis & Miller, 1992; Wayland et al., 1994). The difference in height is not critical, since our analyses focus on the relative locations of the best-exemplar ranges for the series along the VOT continuum, and not on absolute ratings.

(Manuscript received August 25, 2001;
revision accepted for publication October 17, 2002.)