Perception & Psychophysics
1985, 38 (3), 269-276

On the role of visual rate information in phonetic perception

KERRY P. GREEN and JOANNE L. MILLER
Northeastern University, Boston, Massachusetts

It is well established that listeners process segmentally relevant properties of the speech signal in relation to the rate at which the speech was produced. We investigated whether the critical rate information for this effect is limited to the auditory modality or, alternatively, whether visual rate information provided by the talker's face also plays a role. Audio-visual syllables were created by pairing tokens from a moderate-rate, auditory /bi/-/pi/ series with visual tokens of /bi/ or /pi/ produced at a faster or slower rate of speech; these visual tokens provided information about speaking rate, but could not themselves be identified correctly as /bi/ or /pi/. Each audio-visual pairing produced the phenomenal experience of a single, unified syllable, /bi/ or /pi/, spoken at a single rate of speech. The change in visual rate information across the syllables influenced the judged rate of the audio-visual syllables and, more importantly, affected their identification as /bi/ or /pi/. These results indicate that visual information about speaking rate is relevant to the perception of voicing and, more generally, suggest that the mechanisms underlying rate-dependent speech processing have a bimodal (or amodal) component.

In normal conversation, speakers often exhibit large and frequent changes in the rate at which they talk. Although most of this change in overall rate is due to variation in the number and duration of pauses in the utterance (Goldman-Eisler, 1968; Grosjean & Deschamps, 1975), it has recently been shown that there is also significant variation in the rate at which the speech itself is produced (Miller, Grosjean, & Lomanto, 1984). In effect, a change in overall speaking rate alters the temporal fine structure of the speech signal. This poses a potential problem for speech perception in that many of the acoustic properties that specify the identity of phonetic segments are themselves temporal in nature and are susceptible to modification as talkers change their rates of speech.

Consider, as an example, the case of voicing in syllable-initial stop consonants. It is well established that a major property distinguishing voiced from voiceless stop consonants is voice-onset time (VOT). In acoustic terms, VOT can be defined as the interval between the abrupt increase in acoustic energy corresponding to the release of the consonant and the onset of quasi-periodic energy corresponding to the onset of vocal fold vibration. Voiced consonants are typically produced with shorter VOT values than voiceless consonants (Lisker & Abramson, 1964), and numerous perceptual experiments have shown that listeners use this property when identifying consonants as voiced or voiceless—stimuli with shorter VOT values are heard as voiced, whereas those with longer VOT values are heard as voiceless (e.g., Lisker & Abramson, 1970).

The complication introduced by speaking rate is that as rate is altered, so too is the distribution of VOT values for syllable-initial stop consonants. Summerfield (1975) has reported that as speaking rate becomes slower and syllable duration becomes longer, the VOT values of voiced stops remain relatively constant, but those of voiceless stops become systematically longer (cf. Diehl, Souther, & Convis, 1980). In a study currently in progress, we have confirmed this observation for the particular syllables that are the focus of the present paper, /bi/ and /pi/. The important issue for theories of perception is whether listeners take account of this acoustic change when using VOT to identify consonants as voiced or voiceless; that is to say, whether they treat VOT in a rate-dependent manner, rather than an absolute manner.

In a recent series of studies, Summerfield (1981) provided strong evidence for the existence of such rate-dependent processing: As the duration of syllables beginning with /b/. versus/p/ was lengthened, the VOT value that perceptually differentiated /b/ from /p/ also became longer (see also Miller, Dexter, & Pickard, 1984). Moreover, a shift in perceptual criterion as a function of rate is not limited to the case of voicing as specified by VOT, but has been reported for a variety of segmental contrasts specified by a number of temporal parameters (see Miller, 1981, for a review).

The purpose of the present investigation was to explore further the nature of the listener's adjustment for changes in speaking rate during phonetic perception. Our focus

This research was conducted as part of a doctoral dissertation at Northeastern University by the first author under the direction of the second author, and was supported in part by NIH Grant NS 14394 and NIH RCDA NS 00661 to J. L. Miller, and NIH BRSG RR 07143 to Northeastern University. We thank Peter D. Eimas for his valuable comments on an earlier version of the manuscript and Lisa Ventura and Nancy Franey for their assistance in data collection. Some of these data were presented at the fall meeting of the Acoustical Society of America, Minneapolis, 1984. Requests for reprints should be addressed to Kerry P. Green, who is now at the Child Development and Mental Retardation Center, University of Washington, WJ-10, Seattle, WA 98195.

Copyright 1985 Psychonomic Society, Inc.

concerned the particular rate information to which the listener adjusts: We asked whether the critical rate information is limited to the auditory domain, or whether visual rate information provided by the talker's face also plays a role.

Our research was motivated in large part by the recent surge of interest in the role of visual information in segmental perception. It has been known for some time that watching a talker speak can enhance speech recognition if the acoustic signal itself is degraded in some fashion, for example, by presenting it in noise (e.g., Dodd, 1977; Erber, 1969). Recently, however, it has been argued that the visual information provided by the talker's face may play more than a compensatory role in identifying the segmental structure of speech. In a classic paper, McGurk and MacDonald (1976) demonstrated that the presentation of an undistorted auditory syllable in synchrony with a different visual syllable often led to the percept of a single syllable that was a joint function of the auditory and visual information. To take a recent example of the general phenomenon, Massaro and Cohen (1983) found that when members of an auditory continuum ranging from /ba/ to /da/ were paired with visual tokens of a talker saying either /ba/ or /da/, the visual information had a substantial influence on the observer's identification of the syllables (cf. Summerfield, 1979). What is especially interesting is that in this situation subjects are typically unaware of the bimodal discrepancy inherent in the stimuli; the phenomenal experience is that of a single, unified syllable. The cross-modality effect has been taken as evidence that the speech processing system normally functions to integrate available segmental information from the two modalities and, moreover, that the output of this process is an amodal, phonetic representation (for further discussion, see Liberman, 1982; Summerfield, 1979).

In the present research, we asked whether visual information for speaking rate, like visual information for segmental identity per se, plays a role in phonetic perception. Our basic strategy was to pair auditory tokens from a /bi/-/pi/ continuum that was moderate in rate of speech with a video display of a talker producing /bi/ and /pi/ at faster and slower rates of speech. The phenomenal experience that resulted from any single audio-visual pairing was that of a single syllable, /bi/ or /pi/, produced at a single rate of speech. Two predictions were tested. First, if listeners integrate information about speaking rate from the two modalities, then the change in rate of the visual syllables should influence the judged rate of the audio-visual syllables. Second, and most important, if the visual rate information plays a role in phonetic perception, then the change in rate of the visual syllables should also affect the identification of the audio-visual syllables as /bi/ or /pi/.

Three experiments were conducted in all. The first two, which were preliminary in nature, focused separately on the auditory and visual tokens that would be paired to create the audio-visual syllables. The third, and main, experiment assessed perception of the audio-visual syllables themselves.

EXPERIMENT 1

This experiment focused on the moderate-rate auditory /bi/-/pi/ syllables that were to be paired with the visual syllables in Experiment 3. Its primary purpose was to confirm that a change in speaking rate that was specified auditorily (by a change in syllable duration) did influence identification of these particular syllables, as expected on the basis of previous results (e.g., Summerfield, 1981). This influence would be seen as a shift in location of the /b/-/p/ category boundary toward shorter VOT values as syllable duration was decreased, and toward longer VOT values as syllable duration was increased. A secondary purpose was to confirm that a change in syllable duration also influenced the perceived speaking rate of the syllables.

Method

Subjects. The subjects were 10 undergraduate students who were paid for their participation in the experiment. None reported any history of a speech or hearing disorder.

Stimuli. The stimuli consisted of three auditory /bi/-/pi/ series. The stimuli within each series varied in VOT value from 7 to 57 msec; the series differed from each other in the overall duration of the syllables. The stimuli were computer-edited versions of natural speech, created as follows:

The first step was to construct a single /bi/-/pi/ series whose stimuli varied in VOT. A male native speaker of English was recorded saying several tokens of the syllables /bi/ and /pi/ in isolation. These were recorded in a small, quiet room using a Revox B77 tape recorder and an AKG D200E microphone. The syllables were digitized at a 20-kHz sampling rate with a speech processing facility implemented on a PDP-11/44 computer. One /bi/ and one /pi/ whose overall syllable durations were nearly identical (302 and 305 msec, respectively) were chosen for further processing. Specifically, a cross-splicing technique (cf. Ganong, 1980) was used to produce a 13-member series that varied in VOT value from 7 msec, the VOT value of the original /bi/, to 57 msec, the VOT value of the original /pi/. This technique involved deleting successively larger acoustic segments from /bi/, beginning at the release of the consonant, and replacing these with equally long acoustic segments from /pi/, again beginning at the release of the consonant. The durations of the successively longer acoustic segments, and hence by definition the VOT values of the stimuli, were 7, 14, 19, 21, 26, 29, 33, 37, 40, 43, 47, 50, and 57 msec. All cuts in the /bi/ were made at a zero-crossing and those at 7, 14, 21, 29, 37, 43, 50, and 57 msec were also located at the beginning of a pitch period. There was no audible indication (e.g., the presence of a click) that any of the syllables had been edited; all sounded like good tokens of natural speech.

The second step involved creating two additional /bi/-/pi/ series that differed from the first in overall syllable duration. This was accomplished by simply deleting the final 92 or 141 msec of vowel from each stimulus in the original series. This process resulted in three different series that were identical except for the stimulus endings. The three series, with overall syllable durations of 161, 210, and 302 msec, are labeled (on the basis of relative rate) as the fast, moderate, and slow series, respectively.

Eleven different randomized orders of the 39 stimuli (3 series × 13 stimuli each) were recorded onto one audio channel of a ¾-in. videotape (Sony VO 2611 videocassette recorder) for presentation to the subjects. There was a 4-sec interval between stimuli within an order and a 10-sec interval between orders.

Procedure. Each subject participated in two 40-min sessions conducted on different days. During each session, the subject was presented with the 11 stimulus orders and was asked to perform two tasks. The first, and primary, task was to identify each syllable as /bi/ or /pi/, indicating the response by writing B or P on a response sheet. The second task was to judge the speaking rate of the syllable, using a 10-point rating scale on which low numbers corresponded to slow speaking rates and high numbers corresponded to fast speaking rates. The subjects were instructed to judge the rate of each syllable in relation to the range of rates that they encountered in the experiment, and to indicate their response by circling a number between 1 and 10 on the response sheet. The first test order of each session was considered practice, and the data from it were not included in any analyses. Thus, for analysis purposes, each subject contributed a total of 20 responses to each stimulus, for each task.

The subjects were tested individually in a small, dimly lit, quiet room. The videotape was played on a ¾-in. videocassette recorder (Sony VO 2611), with the audio signal presented over a loudspeaker (JBL 6001) hidden behind a cardboard panel which enclosed a 12-in. color monitor (Sony CVM 1250). It seemed as if the sound were coming from the video monitor itself, even though the monitor was turned off throughout this experiment. The subject sat at a small table about 4 ft in front of the monitor. The stimuli were presented at a comfortable listening level, approximately 74 dB SPL for the peak intensity of the vowel.

Results and Discussion

The first question to consider is whether a change in syllable duration produced a change in judged speaking rate. To answer this, we calculated the mean judgment of rate for each of the 13 stimuli in each of the three series, for each subject. The group functions are displayed in Figure 1. It is apparent that the judged rate of the stimuli varied systematically as a function of syllable duration: the shorter the syllable, the higher the judged speaking rate. To obtain a summary measure of the effect, we computed, for each subject, the average rate judgment across the 13 stimuli in each series. Collapsing across the 10 subjects, these mean rate judgments were 7.38, 4.91, and 4.01 for the 161-, 210-, and 302-msec series, respectively. A one-way repeated measures analysis of variance on the mean rate judgments (collapsed across the 13 sitmuli within each series) indicated that the effect of syllable duration was highly reliable [F(2,18) = 55.96, p < .0001]. Post hoc analyses revealed further that all three series differed from each other: the judged rate of the 210-msec series was reliably lower than that of the 161-msec series (p < .001), and reliably higher than that of the 302-msec series (p < .01).[1]

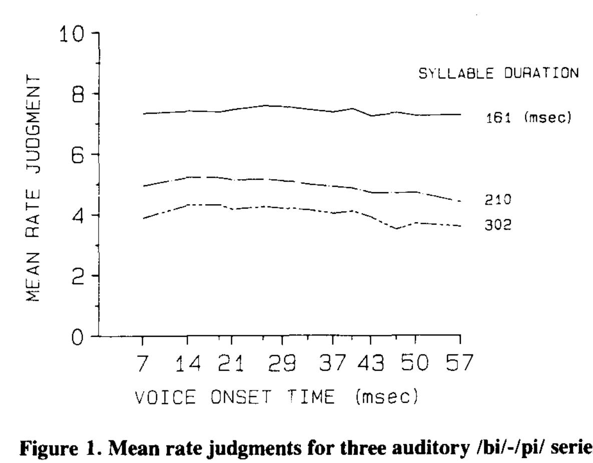

Figure 1. Mean rate judgments for three auditory /bi/-/pi/ series varying in syllable duration (Experiment 1).

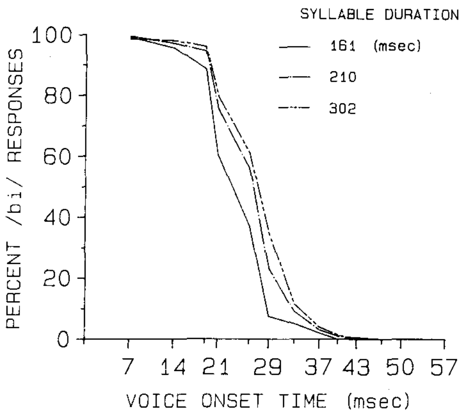

Figure 2. Percentage of /bi/ responses for three auditory /bi/-/pi/ series varying in syllable duration (Experiment 1).

The next question to consider is whether the change in syllable duration influenced the identification of the syllables as /bi/ or /pi/. For each subject, we calculated the percentage of /bi/ responses for each stimulus in each of the three different series. The group identification functions are presented in Figure 2. It is clear that syllable duration influenced identification: as the duration increased, the identification function shifted toward longer VOT values, exactly as predicted. To assess the reliability of this effect, we calculated the location of the individual /b/-/p/ category boundaries for each series. This was done by fitting a linear regression line to the data in the boundary region of the identification function, and taking as the category boundary that VOT value that corresponded to 50% /bi/ responses. The mean boundary values of VOT were 23.55, 25.91, and 26.9 msec for the 161-, 210-, and 302-msec series, respectively. A one-way repeated measures analysis of variance on the individual boundary scores revealed a highly reliable effect of syllabel duration [F(2,18) = 24.33, p < .0001]. Post hoc analyses indicated further that the category boundaries for all three series differed from each other: the boundary value for the 210-msec series was reliably longer than that for the 161-msec series (p < .005, with all 10 subjects showing the effect), and reliably shorter than that for the 302-msec series (p < .05, with 8 of the 10 subjects showing the effect).

The results from this experiment demonstrate that a change in syllable duration influences the judged speaking rate and, of primary importance, the identification of syllables from a /bi/-/pi/ series. In line with previous results (cf. Summerfield, 1981), a decrease in syllable duration from 210 to 161 msec produced a shift in the /b/-/p/ boundary toward shorter VOT values, whereas an increase in syllable duration had precisely the opposite effect. Thus, a change in rate that is specified auditorily (by a change in syllable duration) influences identification of the moderate-rate 210-msec /bi/-/pi/ syllables, making them appropriate to pair with the visual syllables in Experiment 3.

EXPERIMENT 2

The second experiment focused on the fast and slow visual tokens of /bi/ and /pi/ that were to be paired with the auditory 210-msec /bi/-/pi/ syllables in Experiment 3. Its purpose was twofold. First, it was essential to ensure that subjects did perceive the difference in rate between the fast and slow visual tokens. Second, it was critical to demonstrate that subjects could not identify the syllables as /bi/ or /pi/ at better than chance level. If these two things were found to be true, any influence of the visual syllables in Experiment 3 could safely be interpreted as resulting from the visual rate information provided by the syllables, and not from visual information for segmental identity per se. Although previous research has shown that visual information from the face is insufficient for determining the voicing value of consonants (e.g., Binnie, Montgomery, & Jackson, 1974; and see MacDonald & McGurk, 1978), we wanted to establish that this was indeed the case with our specific visual tokens, /bi/ and /pi/.

Method

Subjects. The subjects were 10 undergraduate students none of whom had participated in Experiment 1. All subjects were paid for their participation. None of the subjects reported any history of a speech or hearing disorder and all had normal or corrected-to-normal vision.

Stimuli. The stimuli consisted of fast and slow versions of the syllables /bi/ and /pi/, presented on videotape without sound. Tape preparation involved two steps.

First, the talker who had provided the auditory syllables for Experiment 1 was asked to say a series of fast and slow /bi/s and /pi/s while being video- and audiotaped. The recording apparatus included a Sony DXC 1640 color camera, a Sony VO 2611 ¾-in. color videocassette recorder, and an AKG D200E microphone. The talker was seated in front of a white screen illuminated by two 200-W incandescent bulbs. The camera was centered on the talker's face, which filled about two-thirds of the video frame. Four blocks of 52 trials were recorded. Each block contained 13 fast and 13 slow /bi/s and the same number of fast and slow /pi/s. The four types of syllables occurred in random order, according to a predetermined sequence. On each trial, an assistant cued the talker, by pointing on a card located directly below the camera lens, as to the syllable (/bi/ or /pi/) and rate (fast or slow) he was to produce. The speaker then produced the indicated syllable in time with the next flash of a light located just above the camera lens. The light flashed every 4 sec and served to pace the trials. Note that with this procedure the rate of the individual syllables (fast or slow) varies across trials, but the rate at which the trials occur remains constant. An estimate of the duration of the visual syllables was obtained by measuring each syllable in one block of 52 trials. The mean durations of the fast and slow syllables were 162.0 (SD = 20.4) and 904.0 (SD = 113.0) msec, respectively.

Second, the actual test videotape was prepared by copying the video portion of these four blocks of trials onto another videotape. Two of the blocks were copied twice and the other two were copied once, for a total of six blocks in all. The order in which the blocks were copied was random, with the restriction that two identical blocks could not occur successively. There was a break of approximately 30 sec between blocks. The final test videotape thus consisted of six blocks of 52 trials each, for a total of 78 tokens of each syllable (/bi/ or /pi/) × rate (fast or slow) combination.

Procedure. Each subject participated in a single half-hour session. As in the first experiment, two tasks were required on each trial. First, the subject was to identify the syllable as /bi/ or /pi/, indicating the choice by saying /bi/ or /pi/ out loud. Second, the subject was to judge the speaking rate of the syllable using the 10-point rating scale and, again, to indicate the choice by saying the number out loud. The experimenter recorded the subject's identifications and rate judgments on a response sheet. This procedure allowed the subject to maintain eye gaze on the video monitor at all times.

At the start of the session the subject was presented with 26 practice trials (the first half of one block of trials); the data from these were not included in the analyses. Following the practice trials, the subject was presented with the test videotape. The testing was conducted in the same room, with the same apparatus, as in the first experiment. In the current experiment, however, the video monitor was turned on in order to display the visual tokens, and there was no audio signal.

Results and Discussion

Consider first how the subjects judged the speaking rate of the visual syllables. For each subject, we calculated the mean judged rate of the fast and the slow syllables (collapsed across /bi/s and /pi/s). Averaged across all 10 subjects, these values are 8.23 and 2.44 for the fast and slow tokens, respectively. A paired t test indicated that this difference was highly reliable [$t(9) = 16.23$, $p < .0005$, one-tailed]. Thus the subjects were able to distinguish the rate of the syllables from the visual information alone.

Next consider whether the subjects were also able to identify the syllables as /bi/ or /pi/ on the basis of the visual information. Averaged across subjects, the mean number of correct responses (out of 312 trials) was 159.6, or 51.2%, which was not reliably different from chance performance [$z = .35$, $p > .10$]. Furthermore, separate analyses on the two rates showed that consonant identification was at chance level for both the fast syllables (52.0% correct) and the slow syllables (50.3% correct) [$z \leq .42$, $p > .10$ in each case]. Thus, in accord with previous studies, subjects could not perceive the voicing distinction between /bi/ and /pi/ on the basis of visual information alone. Finally, we tallied the number of /bi/ (as opposed to /pi/) responses separately for the fast and the slow syllables. No bias was found: Averaged across subjects, the mean number of /bi/ responses was 82.7 (53%) for the fast syllables and 86.5 (55%) for the slow syllables, and a paired t test indicated that this difference was not reliable [$t(9) = 1.02$, $p > .10$, two-tailed].

In summary, the results of this experiment established that for the set of /bi/s and /pi/s we recorded, a video display of the talker's face (without sound) provides sufficient information for subjects to discern the rate of the syllables as fast or slow, but not to identify them correctly as /bi/ or /pi/. Thus these visual syllables are appropriate to pair with the audio syllables in Experiment 3; they provide information about speaking rate, but not about segmental identity.

EXPERIMENT 3

In this experiment we assessed perception of audio-visual syllables that were created by pairing tokens from the moderate-rate 210-msec auditory /bi/-/pi/ series (Experiment 1) with the fast and slow visual tokens of /bi/ and /pi/ (Experiment 2). Two questions were addressed: (1) Does the change in rate of the visual syllables from fast to slow influence the judged speaking rate of the audio-visual syllables? (2) Does the change in rate of the visual syllables influence the identification of the audio-visual syllables as /bi/ or /pi/?

Method

Subjects. The subjects were 10 undergraduate students none of whom had participated in the previous experiments. All subjects were paid for their participation. None of the subjects reported any history of a speech or hearing disorder and all had normal or corrected-to-normal vision.

Stimuli. The stimuli for this experiment were audio-visual syllables that were created by pairing auditory and visual tokens of /bi/ and /pi/ as follows. The auditory syllables were the 13 members of the moderate-rate, 210-msec /bi/-/pi/ series tested in the first experiment. The visual syllables consisted of the four original blocks (52 stimuli each) of fast and slow /bi/s and /pi/s tested in the second experiment. For each block of visual syllables, each member of the auditory series was paired once with each type of visual syllable—fast /bi/, fast /pi/, slow /bi/, and slow /pi/.

The procedure used to create the audio-visual pairings was as follows. The original 13 syllables from the 210-msec /bi/-/pi/ series were digitally transferred and stored on a PDP-11/23 computer. The videotape of the four blocks of 52 trials each was played, and the output of the audio channel that contained the original auditory productions of these fast and slow syllables was fed into a Schmitt trigger on the computer. When the Schmitt trigger detected the onset of the original auditory syllable, it triggered the computer to output one of the members of the 210-msec auditory series, according to a predetermined protocol. This syllable was output at a 20-kHz sampling rate and fed to the input of the second audio channel on the videocassette recorder. This procedure enabled us to dub the 210-msec auditory syllables onto the fast and slow visual syllables with high accuracy. The discrepancy between the onset of the dubbed audio syllable and the original audio syllable was measured for the first 10 trials of one of the stimulus blocks and was found to be no greater than 8 msec; this is well below the difference of approximately 80 msec that is required by subjects to detect audio-visual asychrony (McGrath and Summerfield, 1985).

The final test videotape was created by copying each of the 4 blocks of trials onto another videotape three times, for a total of 12 blocks, or 624 trials. The order of the blocks on the final videotape was randomized, with the restriction that the same block could not occur twice in succession. There was a break of approximately 30 sec between blocks on the tape.

Procedure. Each subject participated in two half-hour sessions, conducted on different days. As in the previous experiments, the subject was asked to perform two tasks on each trial: identify the syllable as /bi/ or /pi/ and judge the speaking rate of the syllable using the 10 point rating scale. The subject responded verbally on each trial, saying /bi/ or /pi/ for the identification task and a number from 1 to 10 for the rate judgment task. An experimenter wrote down these responses on an answer sheet.

At the start of each session the subject was presented with 26 practice trials (the first half of one block of trials); responses to these trials were not included in the analyses. After the practice trials, the subject was presented with six experimental blocks—Blocks 1 through 6 in the first session and Blocks 7 through 12 in the second session. Thus 624 trials were presented in all. Collapsing across visual /bi/s and /pi/s, this yields 24 trials for each auditory stimulus (13 stimuli varying in VOT) by visual rate (fast or slow) combination.

Testing took place in the same room as in the previous experiments, and the same equipment was used for stimulus presentation. In this experiment, however, both the video monitor and the loudspeaker were turned on to present the audio-visual syllables. As in Experiment 1, the audio signal was presented at approximately 74 dB SPL, measured for the peak intensity of the vowel. Pilot testing with naive subjects corroborated our own impression that with our experimental setup, each trial was perceived as an instance of a talker saying a single, unified, /bi/ or /pi/.[2]

Results and Discussion

Consider first whether the change in visual rate information from fast to slow affected the rate judgments of the audio-visual syllables. We calculated, for each subject, the mean rate judgment for each member of the moderate-rate 210-msec auditory series when paired with the fast visual syllables and when paired with the slow visual syllables. These data, averaged across the 10 subjects, are presented in Figure 3. Clearly, the visual rate information influenced the judgments. Collapsed across the 13 stimuli in the /bi/-/pi/ series (and across the 10 subjects), the mean rate judgments were 6.20 and 3.83 for the fast and slow visual pairings, respectively. A paired t test indicated that this difference was highly reliable [t(9) = 4.13, p < .001, one-tailed]. It is of interest to note that these values are less extreme than the rate judgments obtained in Experiment 2, in which only the visual portion of the audio-visual syllables was presented (ratings of 8.23 and 2.44 for the fast and slow visual syllables, respectively).[3] This suggests that the subjects were not basing their judgments solely on the visual change in rate, but were using both auditory and visual rate information.

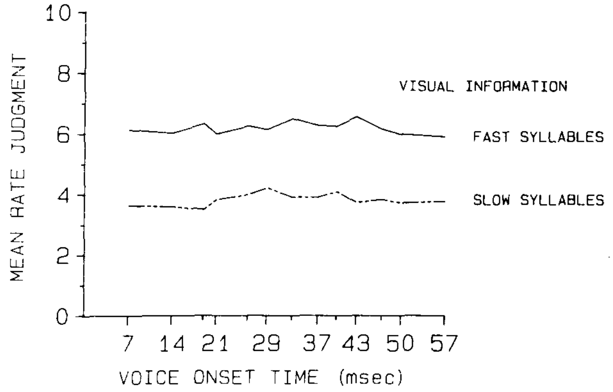

Figure 3. Mean rate judgments for syllables from a moderate-rate auditory /bi/-/pi/ series when paired with fast or slow visual syllables (Experiment 3).

Finally, did the change in rate of the visual syllables influence not only the judged rate of the audio-visual syllables, but their identification as /bi/ or /pi/ as well? To answer this question we computed, for each subject, the percentage of /bi/ responses for each member of the 210-msec series when paired with the fast visual syllables and when paired with the slow visual syllables. The group identification functions for these two conditions are presented in Figure 4. The change in rate of the visual syllables from fast to slow produced a shift in the identification function toward longer VOT values, replicating closely the pattern of results found in Experiment 1, where the rate of the auditory stimulus was itself altered. To obtain a summary measure of the effect of the visual rate information, we calculated the location of the individual /b/-/p/ category boundaries for each visual rate condition, using the same procedure as in Experiment 1. The mean boundary values, averaged across subjects, were 24.06 and 26.34 msec VOT for the fast and slow visual conditions, respectively. A paired t test indicated that the effect, which was shown by all 10 subjects, was highly reliable [t(9) = 5.81, p < .0005, one-tailed]. This finding provides strong evidence that visual rate information plays a role in segmental perception.

An issue that immediately arises concerns the specific conditions under which visual rate information affects phonetic perception. We have argued in earlier papers that the use of the auditory rate information during phonetic perception may in fact be obligatory, that is, that listeners cannot process speech without taking account of the rate at which it was produced (Miller, Dexter, & Pickard, 1984; Miller, Green, & Schermer, 1984). Is the same true of visual rate information? In a first attempt to address this issue, we conducted a modified version of the third experiment on a new group of subjects. Recall that in the third experiment, subjects were required to make two responses on each trial. They identified each audio-visual syllable as /bi/ or /pi/ and they judged its rate. Thus the subjects were explicitly attending to the change in rate of the syllables across trials. In the replication experiment, we asked only for identification responses; the fact that the syllables were produced at different speaking rates was never mentioned in the instructions. Nevertheless, we obtained precisely the same pattern of identification results. The mean /b/-/p/ category boundary for the fast condition, 22.01 msec, was reliably shorter than that for the slow condition, 22.95 msec [t(9) = 1.87, p < .05, one-tailed], with 7 of the 10 subjects showing the effect. Although we cannot say that the use of visual rate information is obligatory, it is at least the case that this information influences segmental identity even when attention is not explicitly directed toward changes in rate.

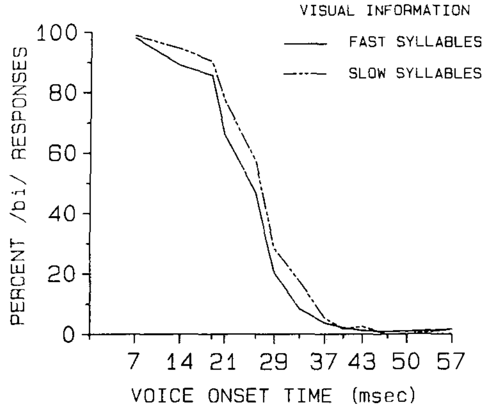

Figure 4. Percentage of /bi/ responses for syllables from a moderate-rate auditory /bi/-/pi/ series when paired with fast or slow visual syllables (Experiment 3).

GENERAL DISCUSSION

The current investigation was based on two findings in the literature. First, there is considerable evidence that listeners process segmentally relevant temporal properties of the speech signal in relation to the rate at which the speech is produced (cf. Miller, 1981). Second, it has been demonstrated that visual information about segmental identity provided by a talker's face can affect identification of the phonetic structure of speech (e.g., McGurk & MacDonald, 1976). We asked whether visual information about speaking rate also plays a role in deriving the phonetic structure of an utterance.

Our strategy was to assess perception of audio-visual tokens of /bi/ and /pi/ that varied in rate. The /b/-/p/ segmental distinction was specified acoustically by a change in VOT, whereas the rate difference between fast and slow was specified visually by a change in the articulatory movements evident on the talker's face. The results were straightforward. The visual change in rate not only influenced the judged rate of the audio-visual syllables, but affected their identification as /bi/ or /pi/ as well: When the visual information changed from specifying a fast to a slow speaking rate, there was a concomitant shift in the location of the /b/-/p/ category boundary toward longer VOT values. This pattern of results replicates the effect of actually altering the rate of the auditory syllables themselves by changing syllable duration, and indicates that visual information about speaking rate is relevant, if not necessary, to the perception of voicing.

This finding raises a number of issues germane to models of phonetic perception and, in particular, to models of the rate effect. Although no detailed account of the mechanisms underlying the rate effect has been profferred in the literature, two general classes of models have been considered. (For discussions of the relative merits and inadequacies of these models see Miller, Aibel, & Green, 1984; Port & Dalby, 1982; Summerfield, 1981; cf. Fowler, 1977, 1980). In the first class of models, labeled extrinsic timing accounts, the listener extracts two types of information during the course of speech processing—information about segmental structure per se and information about rate. The rate information is, in effect, extrinsic to the critical segmental information. The final decision on segmental identity is made on the basis of the segmental information in conjunction with the rate information. To accommodate the present findings within this type of account, it must be assumed either that the decision process has access to both auditory and visual rate information and somehow integrates the two, or, alternatively, that the rate information is itself represented in a metric that is not modality-specific.

A closely related issue that arises within an extrinsic timing framework is whether the rate information for which the listener adjusts is properly defined in terms of physical or subjective rate. Consider the current finding. A physical change in rate (specified visually) produced a modification both in the judged speaking rate of the audio-visual syllables and in their identification as /bi/ or /pi/. But was the change in identification actually mediated by the change in subjective rate, or, alternatively, did it derive more directly from a representation of the physical rate information? Although this issue has not been investigated in the case of audio-visual rate information, there are relevant data from a study we conducted recently on auditory syllables (Miller, Aibel, & Green, 1984). This investigation focused on the /b/-/w/ distinction, which, like the /b/-/p/ distinction, is processed in a rate-dependent manner. Our strategy was to use a contrast paradigm to alter the subjective speaking rate of the auditory syllables without altering the physical rate (duration). We then assessed whether this change in subjective rate produced a change in /b/-/w/ identification. The answer was a clear no, indicating that at least in the case of auditory rate information, the rate effect is not mediated by a change in the subjective experience of a syllable's rate. It will be of interest to determine whether the same is true in the case of the audio-visual rate effect.

Models of the second general class, labeled intrinsic timing accounts (cf. Summerfield, 1981), assume that the correct specification of the critical acoustic information for a given phonetic segment remains invariant under any transformation due to a change in rate, and segmental identity depends on this invariant information. Thus there is no need for a separate process that extracts information about speaking rate from the acoustic signal and normalizes for it. To take an example, the critical segmental information for the /b/-/p/ voicing distinction studied in the present investigation would be not VOT per se, but some higher order variable defined over VOT and syllable duration that remained constant as rate (and hence syllable duration) changed. The present findings have an interesting implication for this type of model. To accommodate them, it must be assumed that the critical segmental information that remains invariant under the rate transformation is not defined solely in acoustic terms (or psychoacoustic terms; cf. Pisoni, Carrell, & Gans, 1983), but rather in some metric that can represent segmentally relevant information from both the auditory and visual modalities—perhaps a metric defined in articulatory or phonetic terms (cf. Liberman, 1982; Summerfield, 1979).

Finally, consider the ontogenetic origins of the ability to accommodate for changes in speaking rate during speech processing. It has been shown that young infants possess at least the rudiments of the ability to adjust for rate variation during speech processing: Infants, like adults, process segmentally relevant temporal properties of speech in relation to syllable duration, which provides information about speaking rate (Eimas & Miller, 1980). It has also been shown that young infants are sensitive to correspondences in auditory and visual information for speech events (Kuhl & Meltzoff, 1982; MacKain, Studdert-Kennedy, Spieker, & Stern, 1983). The interesting question raised by our findings is whether for the infant, as for the adult, the mechanisms underlying constancy across rate variation are not limited to the auditory modality, but instead operate in terms of a bimodal (or amodal) specification of the relevant information for speech perception.

REFERENCES

BINNIE, C. A., MONTGOMERY, A. A., & JACKSON, P. M. (1974). Auditory and visual contributions to the perception of consonants. *Journal of Speech & Hearing Research*, **17**, 619-630.

DIEHL, R. L., SOUTHER, A. F., & CONVIS, C. L. (1980). Conditions on rate normalization in speech perception. *Perception & Psychophysics*, **27**, 435-443.

DODD, B. (1977). The role of vision in the perception of speech. *Perception*, **6**, 31-40.

EIMAS, P. D., & MILLER, J. L. (1980). Contextual effects in infant speech perception. *Science*, **209**, 1140-1141.

ERBER, N. P. (1969). Interaction of audition and vision in the recognition of oral speech stimuli. *Journal of Speech & Hearing Research*, **12**, 423-424.

FOWLER, C. A. (1977). *Timing control in speech production*. Bloomington: Indiana University Linguistics Club.

FOWLER, C. A. (1980). Coarticulation and theories of extrinsic timing. *Journal of Phonetics*, **8**, 113-133.

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

GOLDMAN-EISLER, F. (1968). *Psycholinguistics: Experiments in spontaneous speech*. New York: Academic Press.

GROSJEAN, F., & DESCHAMPS, A. (1975). Analyse contrastive des variables temporelles de l'anglais et due francais: Vitesse de parole et variables composants, phenomenes d'hesitation. *Phonetica*, **31**, 144-184.

KIRK, R. E. (1968). *Experimental design: Procedures for the behavioral sciences*. Belmont, CA: Wadsworth.

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

LIBERMAN, A. M. (1982). On finding that speech is special. *American Psychologist*, **37**, 301-323.

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 Conference of Phonetic Sciences* (pp. 563-567). Prague: Academia.

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

MACKAIN, K., STUDDERT-KENNEDY, M., SPIEKER, S., & STERN, D. (1983). Infant intermodal speech perception is a left-hemisphere function. *Science*, **219**, 1347-1349.

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

McGrath, M., & Summerfield, Q. (1985). Intermodal timing relations and audio-visual speech recognition by normal-hearing adults. *Journal of the Acoustical Society of America*, **77**, 678-685.

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

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

Miller, J. L., Aibel, I. L., & Green, K. (1984). On the nature of rate-dependent processing during phonetic perception. *Perception & Psychophysics*, **35**, 5-15.

Miller, J. L., Dexter, E. R., & Pickard, K. A. (1984). Influence of speaking rate and lexical status on word identification. *Journal of the Acoustical Society of America*, **76**, S89.

Miller, J. L., Green, K., & Schermer, T. M. (1984). A distinction between the effects of sentential speaking rate and semantic congruity on word identification. *Perception and Psychophysics*, **36**, 329-337.

Miller, J. L., Grosjean, F., & Lomanto, C. (1984). Articulation rate and its variability in spontaneous speech: A reanalysis and some implications. *Phonetica*, **41**, 208-214.

Pisoni, D. B., Carrell, T. D., & Gans, S. J. (1983). Perception of the duration of rapid spectrum changes in speech and nonspeech signals. *Perception & Psychophysics*, **34**, 314-322.

Port, R. F., & Dalby, J. (1982). Consonant/vowel ratio as a cue for voicing in English. *Perception & Psychophysics*, **32**, 141-152.

Summerfield, A. Q. (1975). Aerodynamics versus mechanics in the control of voicing onset in consonant-vowel syllables (Speech Perception No. 4). Belfast: Queen's University, Department of Psychology.

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

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

NOTES

[1] All post hoc analyses reported in this paper were done with the least significant difference test (Kirk, 1968).

[2] It is interesting that this was true even in the case of the slow visual syllables, since their duration (approximately 900 msec) was substantially longer than that of the auditory syllables with which they were paired (210 msec). The phenomenal experience of these syllables was that of a talker rapidly opening his mouth at syllable onset and then very gradually closing his mouth. The auditory signal seemed to gradually fade away throughout the period of mouth closing—in a sense, a type of perceptual restoration.

[3] Statistical analyses on the mean rate judgments in the two experiments confirmed that these differences were reliable. A two-way analysis of variance indicated a significant effect of speaking rate [$F(1,18) = 145.9$, $p < .0001$]. The effect of experiment was not significant [$F(1,18) = 1.25$, $p > .10$], but there was a reliable rate by experiment interaction [$F(1,18) = 25.63$, $p < .0005$]. Post hoc analyses indicated that the ratings for the fast syllables in Experiment 2 (mean of 8.23) were reliably higher than for the fast syllables in Experiment 3 (mean of 6.20), and that the mean ratings for the slow syllables in Experiment 2 (mean of 2.44) were reliably lower than for the slow syllables in Experiment 3 (mean of 3.83) ($p < .01$ in each case).

(Manuscript received June 24, 1985;
revision accepted for publication September 11, 1985.)