Perception & Psychophysics
2000, 62 (7), 1405-1412

Effects of talker variability on speechreading

DEBORAH A. YAKEL, LAWRENCE D. ROSENBLUM, and MICHELLE A. FORTIER
University of California, Riverside, California

The effects of talker variability on visual speech perception were tested by having subjects speechread sentences from either single-talker or mixed-talker sentence lists. Results revealed that changes in talker from trial to trial decreased speechreading performance. To help determine whether this decrement was due to talker change—and not a change in superficial characteristics of the stimuli—Experiment 2 tested speechreading from visual stimuli whose images were tinted by a single color, or mixed colors. Results revealed that the mixed-color lists did not inhibit speechreading performance relative to the single-color lists. These results are analogous to findings in the auditory speech literature and suggest that, like auditory speech, visual speech operations include a resource-demanding component that is influenced by talker variability.

The relationship between speaker and speech recognition has been explored extensively in the last 40 years. Many theories of auditory speech perception propose that speech input undergoes a normalization process in which talker-specific attributes are extracted and discarded, leaving the phonetic material needed for the perception of speech segments (Halle, 1985; K. Johnson, 1990). The concept that this talker normalization process might also extend to audiovisual speech recognition is implicit in other speech perception theories (Fowler, 1986; Liberman & Mattingly, 1985; McClelland & Elman, 1986). However, no research has addressed this question for visual speech. By borrowing the recent methods used in the auditory speech literature (Mullennix, Pisoni, & Martin, 1989; Sommers, Nygaard, & Pisoni, 1994), the present study examines the degree to which *visual* talker normalization influences speechreading (lipreading).

The issue of visual speech normalization would seem important for two reasons. First, there is accumulating evidence that visual speech perception is an important component of the general speech perception process. While it is clear that speechreading can be useful for the hearing impaired, it is also known that visual speech is used by individuals with good hearing when they are faced with a noisy environment (e.g., MacLeod & Summerfield, 1987), speech with a heavy foreign accent, or speech conveying complicated subject matter (Reisberg, McLean, & Goldfield, 1987). There is also evidence that access to visual speech is necessary for normal speech development (Mills, 1987). Finally, in one experimental context using discrepant audiovisual speech syllables, integration of visual speech was shown to be automatic and mandatory for most subjects (McGurk & MacDonald, 1976).

The second reason visual speech normalization would seem an important issue is that it bears on a general theoretical question in cognitive science. The question of modularity (see, e.g., Fodor, 1983), or whether particular cognitive functions exhibit behavioral and anatomical specialization, has been central to modern cognitive science. Among other characteristics, modules are considered to be informationally encapsulated in that they have access to only the information/processes needed for their particular function. Interestingly, two of the prototypical modules cited by theorists are those for speech/language and face perception (e.g., Ellis, 1989; Fodor, 1983; Liberman & Mattingly, 1985). In this sense, visual speech perception would seem to pose a particularly interesting theoretical problem: Are processes enlisted for visual speech perception associated with speech, face, or both functions? The modular characteristic of information encapsulation would seem to suggest that all language recognition—including visual speech perception—would discard talker-specific facial properties. Testing the influences of talker/face normalization on visual speech perception can help address this question. We now turn to the auditory speech normalization literature in order to borrow conceptual and methodological tools.

One way the effects of auditory speech normalization have been examined is through measuring the processing costs incurred during listening to multiple- versus single-talker stimuli. It is known that vowel and consonant stimuli are easier to identify when spoken by a single talker than when the talker changes from trial to trial (e.g., Strange, Verbrugge, Shankweiler, & Edman, 1976; Verbrugge, Strange, Shankweiler, & Edman, 1976). Analogous multiple-talker effects have been observed for latencies in vowel categorizing and matching (Summerfield and Haggard, 1973). With regard to more complex stimuli,

This research was supported by NSF Grant SBR-9617047 awarded to L.D.R. We gratefully acknowledge the assistance of Elizabeth Alberto, Mike Gordon, Sheila Kirby, Anjani Panchal, Julia Rogenski, and Christi Royster, as well as the helpful comments of two anonymous reviewers and the UCR cognitive science group. D. A.Y. is currently in the Department of Psychology at Orange Coast College. M.A.F. is currently in the Department of Psychology at San Diego State University. Correspondence should be addressed to L. D. Rosenblum, Department of Psychology, University of California, Riverside, CA 92521 (e-mail: rosenblu@ citrus.ucr.edu).

Creelman (1957) found that listeners identify words embedded in noise less accurately when they are spoken by multiple versus single talkers. Mullennix et al. (1989) replicated these findings using a larger set of words both embedded in noise and in the clear. In one experiment, 11 subjects received a list of 68 words produced by one of the talkers, while another 11 subjects received a list of 68 words derived from 15 talkers. Both word lists were presented against varying degrees of white noise. Word identification was more accurate for the single- than for the mixed-talker list. Concerned that the effects might be attributable to the degraded nature of the stimuli, Mullennix et al. conducted a second experiment with undegraded words and measured response latency for a naming task. Performance for both identification and latency measures were significantly worse for the mixed- than for the single-talker list.

Mullennix and his colleagues (1989) offered two explanations for these results. First, the effects of talker variability could be due to speaker normalization processes operating at a very early stage in the acoustic–phonetic analysis. The normalized output would then be passed on to higher level language processes without talker-specific information. On the basis of this account, one would expect processing costs due to talker variability to occur early in speech perception but not during higher level processing.

Alternatively, talker variability may affect performance because talker-specific features are retained for some time, rather than discarded. Retaining such talker-specific features for mixed-talker lists would incur a greater processing cost than it would for single-talker lists. Potentially, talker-specific features from a previous item could produce interference when a subsequent item with different talker-specific features is perceived (Mullennix et al., 1989). From this account, the effects of talker-specific dimensions could appear for higher level functioning.

In fact, other research has supported this latter explanation. There is now evidence that talker-specific information can be retained and that it can act to facilitate speech recognition. For example, Nygaard, Sommers, and Pisoni (1994) trained two groups of subjects to recognize the voices of 10 talkers over a 9-day period. One group was then tested with novel words embedded in noise spoken by unfamiliar talkers while the other group was tested with the novel words spoken by the 10 familiar talkers used during the training phase. The results revealed that familiarity with the talker significantly improved word identification accuracy. There is also evidence that explicit and implicit memory for words is facilitated when study and test items are presented in the same voice (Church & Schacter, 1994; Craik & Kirsner, 1974; Palmeri, Goldinger, & Pisoni, 1993). Finally, research on form-based priming has revealed that priming effects are contingent on prime and target being presented in the same voice (Saldaña & Rosenblum, 1994).

To summarize, a good amount of recent evidence not only supports Mullennix et al.'s (1989) suggestion that talker-specific information can be retained during phonetic processing, but also suggests that talker-specific information can facilitate speech recognition. In this sense, any "normalization" process that might occur would not completely discard talker-specific information. More generally, recent evidence suggests that the functions of phonetic recognition and voice identification are not as independent as once thought (e.g., Halle, 1985; K. Johnson, 1990). In fact, Remez, Fellowes, and Rubin (1997) have proposed that both functions could use similar acoustic primitives—a contention very different from those of traditional theories of speech and speaker perception (e.g., Pollack, Pickett, & Sumby, 1954; Van Lancker, Kreiman, & Emmorey, 1985).

While research on the effects of auditory talker normalization is abundant, there seems to be little analogous research in the visual speech literature. It is known that speakers vary widely in their visible speech movements, which bears on how difficult they are to speechread (Demorest & Bernstein, 1992; Kricos & Lesner, 1982; Montgomery, Walden, & Prosek, 1987). However, it is not known which, if any, talker-specific facial information might be discarded during visual speech perception. Potentially, a visual speech normalization process would strip away phonetically irrelevant information about the face such as eye color, skin tone, and featural information beyond the mouth. The end product of this normalization might be the retention of only phonetically relevant dimensions, including positions and movements of the lips, tongue, and jaw. Presumably, this process of normalization would take some time, so that speechreading from a multiple-talker list would be more difficult than that from a single-speaker list.

The following experiments were designed to examine the effects of visual speech "normalization" by testing the influences of talker variability on speechreading. The experiments borrow from the method of the Mullennix et al. (1989) auditory speech study discussed above. The first experiment compares subjects' speechreading ability from stimuli that are either from a single talker or from multiple talkers. If talker variability has detrimental effects on speechreading, as it does for auditory word recognition, then speechreading accuracy should be worse for subjects in the multiple-talker condition.

EXPERIMENT 1

Method

Subjects. Sixty-two undergraduates at the University of California, Riverside, participated in the experiment and were either paid $10 or were given course credit as part of a requirement of an introductory psychology course. All subjects reported normal or corrected-to-normal vision, good hearing, and were native speakers of English with no prior speechreading experience. Prescreening tests (lasting 10 min) were given to all subjects to ensure a minimal

speechreading ability for subjects used in the experiment. For the prescreening task, 2 talkers (1 male and 1 female) were recorded, each articulating 20 different sentences from the Revised Bamford-Kowal-Bench Sentence Test (BKB; Bench & Bamford, 1979). The recording and editing methods were the same as discussed below. The prescreening tape consisted of two blocks (one for the male talker and one for the female talker), each consisting of two consecutive presentations of 20 different sentences. After the second presentation of each sentence, subjects were asked to repeat whatever words from the sentence they could speechread. A keyword scoring technique was applied (see below). If a subject passed a criterion of 32.5% recognition accuracy of keywords, they were asked to participate in the critical portion of the experiment. (This criterion was set at 1 *SD* below the mean on the basis of other speechreading research [Bernstein, personal communication, 1997] and a pilot study that tested these particular stimuli with 40 subjects.) On the basis of this prescreening criterion, 22 of the original 62 subjects were eliminated from the study.

Stimuli. Ten speakers (5 men and 5 women) were recorded in a fully lit room with no alteration to their faces. The speakers were recorded articulating 100 BKB sentences (Bench & Bamford, 1979; Rosenblum, Johnson, & Saldaña, 1996), four to five times each. BKB sentences are short (5–7 words), simply constructed sentences (e.g., "The football game is over"). BKB sentences are scored using the Loose Keyword Scoring method (Bench & Bamford, 1979), for which a point is given for each of three key words recognized (e.g., *football, game, over*), with morphological errors permitted. None of the sentences or speakers used in the prescreening test were used in the critical stimuli.

The speakers were instructed to articulate clearly, but not to exaggerate their movements. A Panasonic PVS350 camcorder was used to record the initial videotapes. The speakers were seated 8 ft in front of the camera. The camera was positioned so that the recorded image consisted of the speaker's entire head and neck.

Using a Panasonic 7510 video player and a Panasonic 7500A recorder, 11 presentation tapes were made using the 10 different speakers. Ten single-talker tapes were produced, each consisting of 1 of the 10 talkers articulating the same 90 BKB sentences. Each sentence was actually recorded onto the tape twice, in succession. A 1-sec interstimulus interval (ISI) was used between the first and second presentations of a sentence and a 3-sec ISI was used between different sentences. The sentence pairs were separated into six blocks of 15. A 15-sec ISI separated blocks.

A mixed-talker tape was produced composed of all of the 10 talkers each articulating (a different set of) 9 of the 90 sentences. Again, each sentence was placed on the tape twice, consecutively. The 9 sentences were chosen randomly from the talkers with the constraint that the sentence order was the same as that used in the single-talker tapes, and no 1 talker could be seen producing two different sentences consecutively (ensuring trial-to-trial variation). The ISIs and blocks used for the single-talker tapes were also used for the mixed-talker tape.

Procedure. The 40 subjects who passed the prescreening test were randomly assigned to either the mixed-talker or single-talker conditions. The 20 subjects in the mixed-talker condition were all presented with the mixed-talker tape. The 20 subjects in the single-talker condition saw *one* of the single-talker presentation tapes so that 2 subjects saw each tape (Mullennix et al., 1989). Comparing overall performance across all subjects in the single-talker condition with subjects in the mixed-talker condition ensures that observed differences between conditions would not be based on differences in how easy the talkers were to speechread.

Subjects were seated at a table 5 ft in front of a Panasonic 21-in. video monitor. The only source of illumination was the television monitor and one small light that was focused away from the monitor but shed enough light for the experimenter to record the subject's response.

Subjects were told that they would be seeing a speaking face and that they were to attempt to speechread sentences. They were informed that each sentence would be presented twice, and after the second presentation the experimenter would pause the tape and allow the subject to respond. The subjects were instructed to respond verbally to any words they could recognize. The experimenter recorded the key words that the subject verbalized by circling correct responses on a response sheet. The critical phase of the experiment lasted approximately 30 min for each subject.

Results and Discussion

The data were scored for percent correct identification of the three keywords in each of the BKB sentences (Bench & Bamford, 1979). An average of 55.8% (11.1) keywords were identified in the single-talker condition, and 47.9% (11.75) keywords were identified in the mixed-talker condition. This difference for talker condition was significant at the $p < .05$ level [$F(1,20) = 4.86, p = .034$]. These results suggest that a change in the talker from trial to trial can hinder speechreading performance.

While the decrement imparted by the mixed-talker list might seem relatively small (7.9%), it is well within the range of decrement values observed by Mullennix et al. (1989), based on their *auditory* mixed-list condition. Their set of four experiments, which tested word identification of both degraded and clean auditory stimuli, displayed mean mixed-talker decrement values between 4.4 and 21.0%, with an overall average of 9.3%. Clearly, a number of critical differences exist between the Mullennix et al., and present experiments (e.g., auditory vs. visual material; words vs. sentences; 10 vs. 15 talkers), and this similarity in decrement could be coincidental. Still, it could be that the time course of talker normalization and/or retaining of talker-specific information is similar across auditory and visual speech modalities.

In summary, the results of the first experiment demonstrate that speechreading performance from a variable-talker list is worse than performance from a single-speaker list. However, it is not clear that the detriments produced are a result of *talker* variability per se, or are a result of superficial stimulus variability caused by a general stimulus change from trial to trial. It could be that any trial-to-trial change in the appearance of the stimuli would demand more attention, and thus take away from the attentional resources that could be used for speechreading. A similar consideration was entertained by Sommers et al. (1994) with regard to the auditory mixed-talker effects observed by Mullennix et al. (1989; see also Mullennix & Pisoni, 1990). Sommers et al. tested whether trial-to-trial variation in a phonetically irrelevant, talker-*unrelated* dimension—overall amplitude—would also induce performance decrements relative to single-amplitude lists. They observed that unlike the talker characteristics of speaker identity and speaking rate, variation in overall amplitude *did not* produce decrements in performance. They concluded that variability-induced decrements are not due to the effects of general stimulus uncertainty (and increased attention), but instead, to changes in talker-related dimensions that can have acoustic–

phonetic ramifications. Other researchers have used a similar overall amplitude manipulation to support the same conclusion about memory effects of spoken words (Bradlow, Nygaard, & Pisoni, 1999; Church & Schacter, 1994; Nygaard & Burt, 1996).

On the basis of the analogous consideration of our visual speech effects, a second experiment was designed to examine whether more superficial trial-to-trial changes in our visual stimuli could produce decrements in speechreading performance. For this purpose, 10 different color tints were superimposed onto the images of 2 of our talkers' sentence stimuli. The color dimension was chosen for the control manipulation because, while it is doubtful that it has an influence on visual–phonetic (viseme) perception, image color is known to be relevant to face perception. Color information has been shown to aid recognition of famous faces (Lee & Perrett, 1997), as well as judgments of face gender (Hill, Bruce, & Akamatsu, 1995) and age (Burt & Perrett, 1995). More generally, color information can enhance nonface object identification (see, e.g., Humphreys, Goodale, Jakobson, & Servos, 1994; Price & Humphreys, 1989; but see Biederman & Ju, 1988) as well as motion perception (e.g., Edwards & Badcock, 1996; Gegenfurtner & Hawken, 1996; but see Livingstone & Hubel, 1987, and Ramachandran & Gregory, 1978).

In Experiment 2, subjects were asked to speechread sentences under both single- and mixed-color conditions. This procedure enabled the assessment of the effects of talker and phonetically irrelevant visual stimulus variability. In addition, a second group of subjects were asked to speechread from both single- and multiple-talker lists. Beyond providing a comparison for the single/multiple color tint conditions, the latter group provides a test of whether the multiple-talker effects observed in Experiment 1 would replicate under a within-subjects design.

EXPERIMENT 2

Method

Subjects. Eighty-six undergraduates at the University of California, Riverside, were given credit as part of a requirement of an introductory psychology course. All reported normal or corrected-to-normal vision, good hearing, and were native speakers of English with no reported speechreading experience. Subjects ranged in age from 17 to 26 years. Sixty subjects passed the prescreening criterion used for Experiment 1 and were used in the main experiment.

Stimuli. The 11 stimulus presentation tapes used in Experiment 1 were also used in Experiment 2. In addition, 22 new tapes were made using the stimuli from two of the single-talker tapes of Experiment 1. Two sets of color tapes were produced: one set of 11 was derived from the talker that provided the best mean speechreading scores in Experiment 1 (65.7 correct, as calculated across the 20 subjects in the multiple-speaker condition); the other set was derived from the speaker that provided the worst mean speechreading scores in Experiment 1 (35.3% correct). These 2 speakers were chosen for the color control condition to test whether any differences observed between single/mixed speaker and color conditions were a function of overall ease of speechreading. For each of these 2 speakers, 90 BKB sentences were captured onto a computer using the software program Adobe Premiere. To make the 10 single-color tapes, the Adobe Premiere program was used to apply 10 different color tint filters (black, brown, olive, emerald, red, royal blue, aqua, purple, light pink, and hot pink) to all 90 sentences. Tint values were established in RGB color space as various proportions of red, green, and blue values each ranging from 0 to 255 (see Lee & Perrett, 1997). These values for each tint are listed in Table 1. The tints were applied to the entire picture frame (including face) but were transparent enough to allow the speaking face to be seen.

Table 1
RGB Values for the Color Tints Applied to Images in Experiment 2

Color name	R	G	B
Bright pink	255	0	173
Light pink	255	99	163
Purple	253	83	255
Royal blue	93	0	255
Aqua blue	66	208	255
Emerald green	35	255	46
Olive green	83	118	0
Red	255	0	14
Brown	124	44	0
Black	14	4	0

Note—See text for details.

In order to establish that the color tints were easily distinguishable, a pilot experiment was conducted to test tint discrimination. For this pilot experiment, all 10 different tinted versions of the sentence "The football game is over" were used from the poor speaker's sentence set. Multiple instances of these tokens were recorded onto a presentation videotape arranged in 80 total pairs. Forty of these pairs were composed of tokens with the same color tint, while 40 of the pairs were composed of two tokens with different color tints. This second set was composed of one instance (ordering) of all possible combinations of two tints. All 80 pairs were recorded onto a presentation tape in random order with a 1-sec ISI between tokens in each pair, and a 3-sec ISI between pairs. Ten new subjects with (self-reported) good color vision were each paid $5 to judge the stimuli. Subjects were asked to watch each pair of sentences and to judge whether the color tints of each token of a pair were the same or different. They indicated their judgments by circling "same" or "different" on a response sheet. Mean percent correct for this task was 98.1% (with subject means ranging from 96.3% to 100%), and no single-color pair seemed more difficult than others to discriminate. This pilot study shows that the color tints were highly discriminable and were therefore useful for testing the influence of superficial trial-to-trial changes on speechreading.

For the critical speechreading stimuli, the 10 sets of single-color tinted sentences were each recorded onto a presentation tape. The (90) sentences for each color tape were recorded in the same random order and had the same block organization as in Experiment 1. Each sentence was presented twice with a 1-sec ISI and a 3-sec ISI between novel sentence. A 15-sec ISI separated blocks. A mixed-color presentation tape was also made for each of the 2 speakers. To produce the mixed-color presentation tapes, nine different sentences from each of the single-color sets were recorded onto a new tape in the same random order as that of the single color tapes, with the same ISI, blocks, and ordering constraints that were applied to the mixed-talker tape.

Procedure. The room, viewing distance, lighting, and equipment were the same as in Experiment 1. Subjects were randomly assigned to one of two groups. Twenty subjects participated in the single/mixed talker condition and 40 participated in the single/mixed-color condition; 20 were tested with the stimuli derived from the best talker, and 20 were tested with the stimuli derived from the worst talker.

In the single/mixed talker condition, 10 (of the 20) subjects were presented with the first block of 45 sentences from one of the single-

talker presentation tapes used in Experiment 1, followed by the second block of 45 sentences from the mixed-talker presentation tape used in Experiment 1. For the remaining 10 subjects of this group, presentation order was counterbalanced so that the first block of the mixed-talker tape was seen first, followed by the second block of one of the single-talker tapes. As in Experiment 1, each of the single-talker tapes was seen by 2 of the subjects. The instructions and task were identical to those of Experiment 1.

For each of the single/mixed-color conditions, 20 (of the 40) subjects were presented with the first block of 45 sentences from one of the single-color presentation tapes, followed by the second block of 45 sentences from one of the mixed-color presentation tapes. For the remaining 20 subjects of each of these groups, presentation order was counterbalanced so that the first block of one of the mixed-color tapes was seen first, followed by the second block of one of the single-color tapes. Each single-color tape was seen by 2 of the subjects in each group. The instructions and task were identical to those of Experiment 1.

Results and Discussion

As before, the data were analyzed in terms of keyword correct identification accuracy. In the single-talker condition, subjects identified 51.01% (11.89) of the keywords correctly, while in the multiple-speaker condition, 42.06% (8.1) of the keywords were correctly identified. For the best speaker, during the single-color condition, subjects identified 63.74% (11.49) of the keywords correctly, while in the mixed-color condition, 64.15% (13.21) of the keywords were correctly identified. For the worst speaker, subjects correctly identified 36.6% (13.38) of the keywords in the single-color condition and 36.4% (10.44) of the keywords in the multiple-color condition. These mean values are similar to the mean percent correct values for these speakers in Experiment 1 (65.7% and 35.3% for the best and worst speakers, respectively), suggesting that adding the color tints did not have an *overall* affect on speechreading. A three-way analysis of variance (ANOVA) was conducted on the following factors: condition (talker, color best speaker vs. color worst speaker), variability (single vs. mixed), and order (single first vs. mixed first), with alpha = .05. Two significant main effects were found. First, there was a main effect of variability [$F(1,54) = 14.814$, $p = .0003$], with performance in the single speaker/ color conditions (50.25%) better than performance in the multiple speaker/color conditions (47.54%). Second, there was a main effect of condition [$F(2,54) = 32.730$, $p = .0001$], with subjects in the best speaker/ color condition (63.95%) performing better than those in the worst speaker/color condition (46.54%) or the single/multiple-speaker conditions (36.5%). There was no main effect of order [$F(2,54) = .681$, $p = .5103$].

A significant interaction was observed for variability and condition [$F(2,54) = 22.470$, $p = .0001$]. Accordingly, pairwise comparisons were conducted on single- and mixed-list performance for the talker and the two color conditions. There was a significant effect of variability in the talker condition [$F(1,37) = 7.545$, $p = .0092$], with performance in the single-talker condition (51.01%) better than performance in the multiple-speaker condition (42.06%). This replicates the findings of Experiment 1 and demonstrates that talker variability can inhibit speechreading performance. A significant difference was not found for variability in the color condition using the best speaker [$F(1,19) = .011$, $p = .9170$]. Nor was there a significant difference of color variability using the worst speaker [$F(1,19) = .003$, $p = .9537$] (see means above).

The results of the color manipulation suggest that, unlike talker variability, phonetically irrevelant stimulus variability does not produce detrimental effects on speechreading performance. This finding is analogous to the findings of Sommers et al. (1994), who found that the phonetically irrelevant dimension of overall stimulus amplitude variability also had no influence on auditory speech identification. However, conclusions about our color manipulation must be made with some caution since it is difficult to compare across manipulations that vary different stimulus dimensions. As discussed above, it is known that image coloration does influence face perception (Burt & Perrett, 1995; Hill et al., 1995; Lee & Perrett, 1997). Also, our pilot experiment revealed that the color tints used in the present study were highly discriminable. Still, it could be that the relative salience of the color versus speaker dimensions, or the range over which they each varied, was not perceptually equivalent. This issue of relative perceptual salience was also considered by Sommers et al. in discussing their mixed-amplitude versus mixed-talker effects. These authors suggest that future methods that provide for more direct comparisons of perceived similarity (e.g., multidimensional scaling) should be implemented to address this issue. The same suggestion applies to the present set of visual speech stimulus manipulations.

GENERAL DISCUSSION

The results of the present study demonstrate that talker variability produces detrimental effects on visual speech perception. These results suggest that some resource-demanding component of the visual speech process is activated when the talker changes from trial to trial. These results are consistent with results of previous research on auditory speech perception (e.g., Mullennix et al., 1989; Strange et al., 1976; Summerfield & Haggard, 1973; Verbrugge et al., 1976), and in fact show a decrement similar in magnitude to the auditory speech effects.

In discussing the present visual speech results, it would seem useful to turn to explanations offered for the analogous auditory mixed-talker results. As stated, Mullennix et al. (1989) offered two possible ways that talker variability may demand cognitive resources. First, mixed-talker stimuli could invoke a "normalization" process in which some cognitive effort is taken to strip away phonetically irrelevant information. Alternatively, talker-specific features could actually be encoded with phonetic information—a process that could be more resource demanding under mixed-talker than single-talker conditions. As noted, subsequent research in novel word recognition,

explicit and implicit memory, and form-based priming, has supported this latter interpretation (Craik & Kirsner, 1974; Nygaard et al., 1994; Palmeri et al., 1993; Saldaña & Rosenblum, 1994).

More recently, other theories have been offered to explain the speech–speaker contingencies in the auditory speech literature (e.g., Church & Schacter, 1994; Remez et al., 1997; Sheffert & Fowler, 1995). However, for the present visual speech "normalization" results, the two alternative explanations of Mullennix et al. (1989) provide a good starting point for evaluating our findings. First, the performance decrements observed for the mixed-talker lists could be due to a stripping away of phonetically irrelevant visual speech dimensions. Features ranging from skin tone to eye color to visible talker-specific articulatory (idiolectical) style could be discarded to leave relevant visual speech information. In the second account, the stimulus dimensions that are relevant to face recognition could be retained for some period and induce longer term effects on visual speech recognition.

The present experiments do not distinguish between these explanations. Future research—analogous to the work in auditory speech (e.g., Nygaard et al., 1994)—should be conducted to examine this question. However, there currently exists some findings in the visual speech literature that bear on this question.

First, Walker, Bruce, and O'Malley (1995) found that audiovisual speech integration was influenced by the gender congruency between the face and voice depending on whether subjects were *familiar* with the face. Specifically, when the face and voice were incongruent, subjects that were personally familiar with the faces were less susceptible to visual influences on "heard" speech syllables. This finding suggests that, as for auditory speech, talker/face-specific information can be retained to bear on subsequent visual speech perception. More recently, Schweinberger and Soukup (1998) asked subjects to perform speeded classification of two vowels (/i/ and /u/) portrayed in facial photographs. They found that reaction times were faster when subjects were personally familiar with the faces being portrayed (but see Campbell, Brooks, De Haan, & Roberts, 1996). Although these findings were based on classification of just two visual vowels portrayed in static photographs, they are compatible with our findings: To the degree that our single-talker condition provided familiarization with a specific talker, our findings can be interpreted as evidence that familiarity *facilitated* visual speech perception (of speechread sentences). In fact, performance facilitation via speaker familiarization in the single-talker list provides an alternative interpretation to the notion that the multiple-talker lists *inhibited* visual speech. Still, either interpretation allows the possibility that individual talker characteristics play a role in visual speech perception.

There is also some preliminary evidence that visual talker characteristics can be retained for word recognition tasks (Saldaña, Nygaard, & Pisoni, 1996; Sheffert & Fowler, 1995). To date, however, the evidence suggests that these effects might not be completely analogous to those observed for auditory speech (e.g., Craik & Kirsner, 1974; Palmeri et al., 1993). Sheffert and Fowler (1995) used a continuous recognition paradigm to test for talker facilitation of word recognition from audiovisual stimuli. The audiovisual stimuli were dubbed to allow for independent manipulation of auditory and visual talker information. Sheffert and Fowler were able to replicate the previous auditory talker facilitation effects (e.g., Palmeri et al., 1993); they observed that recognition of words was better if the auditory talker was the same across presentations. However, they found no significant recognition facilitation for visual talker. In follow-up experiments, Sheffert and Fowler asked subjects to explicitly judge whether the visual talker was the same across presentations of a given word. They found that subjects could perform this task at better-than-chance levels. They concluded that although visual talker information can be retained along with spoken words, it does not act to facilitate word recognition. They added that potentially, talker-specific auditory and visual information are preserved differently.

In a similar study, Saldaña et al. (1996) also used a continuous recognition method to test for facilitory effects of visual talker information using audiovisual stimuli. In order to force subjects to attend more to the visual speech information, Saldaña et al. embedded the auditory speech in varying degrees of noise. An additional manipulation involved presenting the faces either articulating along with the auditory words or in an unmoving, static configuration. Like Sheffert and Fowler (1995), Saldaña et al. found auditory, but not visual, talker facilitation for word recognition (across speech in noise conditions). Interestingly, however, Saldaña et al. did find some improved recognition performance in dynamic versus static visual conditions when the same talker was used across presentations. They concluded that visual *articulatory* information—versus simple facial information (conveyed statically)—might be encoded in long-term memory.

The research by Sheffert and Fowler (1995) and Saldaña et al. (1996) suggests that visual talker information can be encoded with lexical items, and that this information might take an articulatory form. Still, it is unclear whether this talker-specific information can facilitate recognition, as it can for auditory speech. Future research should be designed to further address this issue, possibly using conditions that further force subjects to rely on visual speech information. Potentially, this could be accomplished with speechreading—as opposed to audiovisual—conditions and/or with hearing-impaired subjects, who naturally rely more on visual speech. Circumstances in which subjects are required to recover visual speech information and/or have more experience speechreading might reveal stronger evidence for visual talker facilitation of recognition memory. This research could also help examine whether the similarity in decrement observed for mixed-speaker stimuli across modalities (see above) is based on similar encoding strategies.

To summarize, the few studies that have examined the effects of talker familiarity on visual speech effects show some evidence that talker-specific information can be retained. Whether this evidence takes the form of inhibition of visual speech influences on auditory syllable identification (Walker et al., 1995), recognition of the visual talker speaking specific words (Sheffert & Fowler, 1995), or some subtle facilitation with articulatory versus general facial information (Saldaña et al., 1996), it suggests that talker-specific information can be retained for longer durations than was observed in the present experiment (less than 1 h). Thus, we expect that of the two interpretations borrowed from Mullenix et al. (1989), the correct interpretation of our single versus mixed visual talker results is the second—that is, visual talker-specific information is retained, thereby incurring a greater processing cost for mixed-talker stimuli. Future research will reveal whether any of the newer theories of auditory speech–talker contingencies (e.g., Church & Schacter, 1994; Remez et al., 1997; Sheffert & Fowler, 1995) are also applicable to visual speech (cf. Rosenblum, Yakel, & Green, 2000).

If future research continues to suggest that talker-specific information is retained for visual speech perception, then the more general question of the relationship between visual speech and face perception (and modularity) will need to be reevaluated. Most of what is currently known about this relationship comes from neuropsychological research (see Ellis, 1989, for a review). Initial evidence showed a dissociation of these functions in prosopagnosics that can speechread, and aphasics that have trouble speechreading but no trouble recognizing faces (Campbell, Landis, & Regard, 1986). Additional evidence has come from normal subjects who often show a left-hemisphere advantage for speechreading (Campbell et al., 1986) and a right-hemisphere advantage for face recognition (e.g., Young, Hay, McWeeney, Ellis, & Barry, 1985; see Ellis, 1989, for a review). However, recent critiques and counterevidence for both the lesion and normal subject research challenge the evidence for a neuropsychological dissociation (e.g., Hillger & Koenig, 1991; Rosenblum & Saldaña, 1998; Sergent, 1982). Thus, it is too early to determine exactly how neuropsychological findings will bear on the visual speech/face perception question.

A few recent behavioral studies are also relevant to this issue. It has long been known that inverting the image of a face makes it disproportionately difficult to recognize (see Valentine, 1988, for a review). This facial inversion effect has been interpreted as support for a specialization of facial processing. However, recent studies show that facial inversion also inhibits speechreading and audiovisual speech integration (Green, 1994; Jordan & Bevan, 1997; Massaro & Cohen, 1996). Furthermore, there is evidence that visual speech perception, like face perception, is hindered by idiosyncratic disruptions of wholistic/configural information through the "Margaret Thatcher effect" (Rosenblum et al., 2000; Thompson, 1980; Valentine & Bruce, 1985). Thus, recent behavioral evidence suggests that the processes of visual speech and face perception might not be completely independent. Research is currently under way in our laboratory to further investigate the behavioral and neuropsychological relationship between these functions (e.g., J. A. Johnson & Rosenblum, 1996; Yakel & Rosenblum, 1996).

The results of the present study provide the first demonstration that the effects of talker variability found in the auditory speech literature also occur in visual speech perception. Potentially, the resource-demanding operations used by listeners to compensate for the variability between different talkers is not limited to auditory speech perception, but also incurs processing costs during visual speech perception. These findings will need to be incorporated into current theoretical accounts of audiovisual speech perception, which have not explicitly addressed the role of "normalization" of visual speech.

REFERENCES

BENCH, J., & BAMFORD, J. (1979). *Speech–hearing tests and the spoken language of hearing impaired children*. London: Academic Press.

BIEDERMAN, I., & JU, G. (1988). Surface versus edge-based determinants of visual recognition. *Cognition*, **20**, 38-64.

BRADLOW, A. R., NYGAARD, L C., & PISONI, D. B. (1999). Effects of talker, rate, and amplitude variation on recognition memory for spoken words. *Perception & Psychophysics*, **61**, 206-219.

BURT, D. M., & PERRETT, D. I. (1995). Perception of age in adult Caucasian male faces: Computer graphic manipulation of shape and colour information. *Proceedings of the Royal Society of London: Series B*, **259**, 137-143.

CAMPBELL, R. T., BROOKS, B., DE HAAN, E., & ROBERTS, T. (1996). Dissoicating face processing skills: Decisions about lip-read speech expression and identity. *Quarterly Journal of Experimental Psychology*, **49A**, 295-314.

CAMPBELL, R. T., LANDIS, T., & REGARD, M. (1986). Face recognition and lip-reading: A neurological dissociation. *Brain*, **109**, 509-521.

CHURCH, B. A., & SCHACTER, D. L. (1994). Perceptual specificity of auditory priming: Implicit memory for voice intonation and fundamental frequency. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **20**, 521-533.

CRAIK, F. I., & KIRSNER, K. (1974). The effect of speaker's voice on word recognition. *Quarterly Journal of Experimental Psychology*, **26**, 274-284.

CREELMAN, C. D. (1957). Case of the unknown talker. *Journal of the Acoustical Society of America*, **29**, 655.

DEMOREST, M. E., & BERNSTEIN, L. E. (1992). Sources of variability in speechreading sentences: A generalizability analysis. *Journal of Speech & Hearing Research*, **35**, 876-891.

EDWARDS, M., & BADCOCK, D. (1996). Global motion perception: Interaction of chromatic and luminance signals. *Vision Research*, **36**, 2423-2431.

ELLIS, H. D. (1989). Processes underlying face recognition. In R. Bruyer (Ed.), *Neuropsychology of face perception and facial expression* (pp. 41-49). Hillsdale, NJ: Erlbaum.

FODOR, J. A. (1983). *Modularity of mind*. Cambridge, MA: MIT Press, Bradford Books.

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

GEGENFURTNER, K. R., & HAWKEN, M. J. (1996). Interaction of motion and color in the visual pathways. *Trends in Neurosciences*, **19**, 394-401.

GREEN, K. P. (1994). The influence of an inverted face on the McGurk effect (abstract). *Journal of the Acoustical Society of America*, **95**, 3014.

HALLE, M. (1985). Speculations about the representation of words in

memory. In V. A. Fromkin (Ed.), *Phonetic linguistics* (pp. 101-104). New York: Academic Press.

HILL, H., BRUCE, V., & AKAMATSU, S. (1995). Perceiving the sex and race of faces: The role of shape and colour. *Proceedings of the Royal Society of London: Series B*, **261**, 367-373.

HILLGER, L. A., & KOENIG, O. (1991). Separable mechanisms in face processing: Evidence from hemispheric specialization [Special issue: Face perception]. *Journal of Cognitive Neuroscience*, **3**, 42-58.

HUMPHREYS, G. K., GOODALE, M. A., JAKOBSON, L. S., & SERVOS, P. (1994). The role of surface information in object recognition: Studies of a visual form agnosic and normal subjects. *Perception*, **23**, 1457-1481.

JOHNSON, J. A., & ROSENBLUM, L. D. (1996). Hemispheric differences in perceiving and integrating dynamic visual speech information. *Journal of the Acoustical Society of America*, **100**, 2570.

JOHNSON, K. (1990). The role of perceived speaker identity in *F*0 normalization of vowels. *Journal of the Acoustical Society of America*, **88**, 642-654.

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

KRICOS, P. B., & LESNER, S. A. (1982, May). Differences in visual intelligibility across talkers. *The Volta Review*, **84**, 219-225.

LEE, K. J., & PERRETT, D. (1997). Presentation-time measures of the effects of manipulations in colour space on discrimination of famous faces. *Perception*, **26**, 733-752.

LIBERMAN, A. M., & MATTINGLY, I. G. (1985). The motor theory of speech perception revised. *Cognition*, **21**, 1-36.

LIVINGSTONE, M. S., & HUBEL, D. H. (1987). Psychophysical evidence for separate channels for the perception of form, color, movement, and depth. *Journal of Neuroscience*, **7**, 3416-3468.

MACLEOD, A., & SUMMERFIELD, Q. (1987). Quantifying the contribution of vision to speech perception in noise. *British Journal of Audiology*, **21**, 131-141.

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

MCCLELLAND, J. L., & ELMAN, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, **18**, 1-86.

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

MILLS, A. E. (1987). The development of phonology in the blind child. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip-reading* (pp. 145-162). Hillsdale, NJ: Erlbaum.

MONTGOMERY, A. A., WALDEN, B. E., & PROSEK, R. A. (1987). Effects of consonantal contest on vowel lipreading. *Journal of Speech & Hearing Research*, **30**, 50-59.

MULLENNIX, J. W., & PISONI, D. B. (1990). *Talker variability and processing dependencies between word and voice* (Tech. Rep. No. 13). Bloomington: Indiana University.

MULLENNIX, J. W., PISONI, D. B., & MARTIN, C. S. (1989). Some effects of talker variability on spoken word recognition. *Journal of the Acoustical Society of America*, **92**, 1085-1099.

NYGAARD, L. C., & BURT, S. A. (1996). Sources of variability as linguistically relevant aspects of speech. *Journal of the Acoustical Society of America*, **100**, 2572.

NYGAARD, L. C., SOMMERS, M. S., & PISONI, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science*, **5**, 42-46.

PALMERI, T. J., GOLDINGER, S. D., & PISONI, D. B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **19**, 309-328.

POLLACK, I., PICKETT, J. M., & SUMBY, W. H. (1954). On the identification of speakers by voice. *Journal of the Acoustical Society of America*, **26**, 403-406.

PRICE, C. J., & HUMPHREYS, G. W. (1989). The effects of surface detail on object categorization and naming. *Quarterly Journal of Experimental Psychology*, **41A**, 797-828.

RAMACHANDRAN, V. S., & GREGORY, R. L. (1978). Does colour provide an input to human motion perception? *Nature*, **275**, 55-56.

REISBERG, D., MCLEAN, J., & GOLDFIELD, A. (1987). Easy to hear but hard to understand: A lipreading advantage with intact auditory stimuli. In B. Dodd & R. Campbell (Eds.), *Hearing by ear and eye: The psychology of lipreading* (pp. 97-113). Hillsdale, NJ: Erlbaum.

REMEZ, R. E., FELLOWES, J. M., & RUBIN, P. E. (1997). Talker identification based on phonetic information. *Journal of Experimental Psychology: Human Perception & Performance*, **23**, 651-666.

ROSENBLUM, L. D., JOHNSON, J. A., & SALDAÑA, H. M. (1996). Visual kinematic information for embellishing speech in noise. *Journal of Speech & Hearing Research*, **39**, 1159-1170.

ROSENBLUM, L. D., & SALDAÑA, H. M. (1998). Time-varying information for visual speech perception. In R. Campbell, B. Dodd, & D. Burnham (Eds.), *Hearing by eye II: Advances in the psychology of speechreading and auditory–visual speech*. Hillsdale, NJ: Erlbaum.

ROSENBLUM, L. D., YAKEL, D. A., & GREEN, K. P. (2000). Face and mouth inversion effects on visual and audiovisual speech perception. *Journal of Experimental Psychology: Human Perception & Performance*, **26**, 806-819.

SALDAÑA, H. M., NYGAARD, L. C., & PISONI, D. B. (1996). Encoding of visual speaker attributes and recognition memory for spoken words. In D. G. Stork & M. E. Hennecke (Eds.), *Speechreading by humans and machines: Models, systems, and applications* (NATO ASI Series F: Computers and Systems Sciences, No. 150). New York: Springer-Verlag.

SALDAÑA, H. M., & ROSENBLUM, L. D. (1994). Voice information in auditory form-based priming. *Journal of the Acoustical Society of America*, **95**, 2870.

SCHWEINBERGER, S. R., & SOUKUP, G. R. (1998). Asymmetric relationships among perceptions of facial identity, emotion, and facial speech. *Journal of Experimental Psychology: Human Perception & Performance*, **24**, 1748-1765.

SERGENT, J. (1982). About face: Left hemisphere involvement in processing physiognomies. *Journal of Experimental Psychology: Human Perception & Performance*, **8**, 1-14.

SHEFFERT, S. M., & FOWLER, C. A. (1995). The effects of voice and visible speaker change on memory for spoken words. *Journal of Memory & Language*, **34**, 665-685.

SOMMERS, M. S., NYGAARD, L. C., & PISONI, D. B. (1994). Stimulus variability and spoken word recognition: I. Effects of variability in speaking rate and overall amplitude. *Journal of the Acoustical Society of America*, **96**, 1314-1324.

STRANGE, W., VERBRUGGE, R. R., SHANKWEILER, D. P., & EDMAN, T. R. (1976). Consonant environment specifies vowel identity. *Journal of the Acoustical Society of America*, **60**, 213-224.

SUMMERFIELD, Q., & HAGGARD, M. P. (1973). Vocal tract normalization and representation. *Brain Language*, **28**, 12-23.

THOMPSON, P. (1980). Margaret Thatcher: A new illusion. *Perception*, **9**, 483-484.

VALENTINE, T. (1988). Upside-down faces: A review of the effect of inversion upon face recognition. *British Journal of Psychology*, **79**, 471-491.

VALENTINE, T., & BRUCE, V. (1985). What's up? The Margaret Thatcher illusion revisited. *Perception*, **14**, 515-516.

VAN LANCKER, D. V., KREIMAN, J., & EMMOREY, K. (1985). Familiar voice recognition: patterns and parameters, Part I: Recognition of backward voices. *Journal of Phonetics*, **13**, 19-38.

VERBRUGGE, R. R., STRANGE, W., SHANKWEILER, D. P., & EDMAN, T. R. (1976). What information enables a listener to map a talker's vowel space? *Journal of the Acoustical Society of America*, **60**, 198-212.

WALKER, S., BRUCE, V., & O'MALLEY, C. (1995). Facial identity and facial speech processing: Familiar faces and voices in the McGurk effect. *Perception & Psychophysics*, **57**, 1124-1133.

YAKEL, D. A., & ROSENBLUM, L. D. (1996). Face identification using visual speech information. *Journal of the Acoustical Society of America*, **100**, 2570.

YOUNG, A. W., HAY, D. C., MCWEENEY, K. H., ELLIS, A. W., & BARRY, C. (1985). Familiarity decisions for faces presented to the left and right cerebral hemispheres. *Brain & Cognition*, **4**, 439-450.

(Manuscript received July 15, 1998;
revision accepted for publication November 12, 1999.)