JOURNAL OF MEMORY AND LANGUAGE 34, 665–685 (1995)

# The Effects of Voice and Visible Speaker Change on Memory for Spoken Words

SONYA M. SHEFFERT AND CAROL A. FOWLER

*Haskins Laboratories and University of Connecticut*

Recent research suggests that voice information is not discarded during word recognition, but is represented in memory and can serve as a retrieval cue for word recognition. The research reported here asks whether other idiosyncratic aspects of an event in which speech occurs are also retained with speech in memory. Four experiments explored the effects of voice and visible speaker change on spoken word recognition. In each experiment, subjects watched a videotape of speakers producing words. When a word was repeated, the visible speaker, his or her voice, and a feature of wearing apparel were either the same or different from the first presentation. Subjects made recognition judgments based on word identity and characteristics of the speaker. The results indicate that the memory subserving spoken word recognition includes detailed information about a talker's voice and face, but that each is preserved differently. © 1995 Academic Press, Inc.

Speech perception takes as its starting point the sensory registration of a highly variable and context-sensitive speech signal. Individual productions of a word vary acoustically because talkers' vocal tracts differ in size and in anatomical detail; because speech rate, emphasis, and style can vary; and because talkers' dialects differ. Despite the variability, however, listeners usually can identify spoken instances of words as particular word types that they have experienced in the past. Our research explores how word recognition is achieved.

It is generally assumed that language users have a "mental lexicon" in which words are stored as abstract types. This assumption, coupled with the fact that spoken words are variable in the ways just described, have led many researchers to propose processes of normalization in which phonetically or lexically significant signal properties are separated from nonlinguistic properties during early perceptual processing (Joos, 1948; Gerstman, 1968; Johnson, 1990; Ladefoged & Broadbent, 1957; Nearey, 1989). Successful access to a lexical entry, therefore, requires that listeners "preprocess" a spoken word to strip away all but the abstract linguistic information represented in lexical entries. The idea is that, for a perceiver to know that the perceptual record used to access the lexicon (henceforth, "the representation of a spoken word that subserves lexical access") matches a lexical entry, the two representations, that of the input and that in the lexicon, must match.

This view predicts that only the word, and not its carrier, is retained in the representations of a word that subserve lexical access. There is some evidence to support this assumption. Jackson and Morton (1984) asked whether word repetition effects (the facilitation in recognition of hearing a word for the second time as compared to the first) were independent of the voice[1]

The research reported here was supported by NICHD Grant HD-01994 to Haskins Laboratories. The authors thank Barbara Church, Douglas Nelson, David Pisoni, Jay Rueckl, and two anonymous reviewers for comments and suggestions. Carol Fowler is also at Yale University. Correspondence and reprint requests should be addressed to Sonya M. Sheffert, Psychology Department, University of Connecticut, Box U-20, Storrs, CT 06268.

[1] In this experiment, and most others that we summarize including our own, when we refer to effects of "voice" preservation we cannot know, in fact, that voice quality, rather than the speaker's dialect, intonation or rate of speech was the operative factor. Ac-



0749-596X/95 $12.00

in which words were spoken. A priming phase of their experiment consisted of animacy judgments of 100 nouns presented in either a male or female voice. The gender of the voice in which the word was spoken during a subsequent identification test was either the same as or different from the voice in the priming phase. Jackson and Morton report equivalent repetition priming effects for test words spoken in either the same or a different voice from the study condition. The lack of effect of a voice change across study and test supports the notion that the input to the lexical system is normalized by abstraction. Repetition effects, they argue, are the result of a temporary lowering of the threshold for "firing" of a "logogen" that represents a given word in the lexicon. When a word is recognized, its logogen fires, and that lowers its threshold for firing to a subsequent occurrence of the same word. Because logogens are abstract, information about the specific visual or acoustic forms of words, and the context in which they occur, is not retained in lexical memory.

However, this conceptualization of the spoken word recognition process may be undermined by several findings showing that idiosyncratic voice information is not discarded during perception. Craik and Kirsner (1974) administered a continuous recognition memory task in which a word could be repeated in one of two voices (male or female), over a maximum lag of 32 trials. They found that performance was enhanced in "same voice" trials. That is, subjects were more accurate at recognizing "old" words (words presented for the second time in the experiment) as such when the same speaker, rather than a different speaker produced the word on the first and second occasions. The fact that preservation of the word-voice pairing across trials facilitated word recognition is identified as an "implicit" effect of memory: subjects were not instructed to encode the speakers' voices, and the word recognition task was performed without intentional recollection of voice information.

Palmeri, Goldinger, and Pisoni (1993) extended this finding by increasing the number of speakers to 20 (and consequently unconfounding the relationship between voice and gender) and lengthening the lag between repetitions. The words were the same across study and test, but half of the test words were spoken in a new voice. They replicated Craik and Kirsner's findings by demonstrating that recognition performance was attenuated by a change in voice between study and test. Furthermore, the recognition advantage for same voice words remained robust even when as many as 64 items intervened between occurrences of a word. The benefit to word recognition in the same voice condition indicates that perceptual details about the carrier of the semantic information were retained in long-term memory.

Unconscious sensitivity to the physical match between repeated items is regarded as evidence for retrieval of information either from a "presemantic representation system" (PRS), in which detailed perceptual records of particular word forms are retained (cf., Schacter, 1992) or from episodic memory, which represents specific instances of events (Feustal, Shiffrin, & Salasoo, 1983; Richardson-Klavehn & Bjork, 1988). Of course, neither of these memory systems has been viewed as including lexical memory. Accordingly, the findings from these studies need not imply that voice is preserved in the representations of words that subserve lexical access. However, other results suggest that it is.

Goldinger (1992) examined implicit memory for spoken word exemplars as a function of the number of speakers in a list (2, 6, or 10), and the time delay between the end of study and the beginning of test (5 min, 1 day or 1 week). The experimental task was to identify words presented in noise by typing their spellings into a computer keyboard, a response that required subjects to

cordingly, we use "voice" as a cover term for acoustic information other than information about the identity of a spoken word.

access the lexical representation of a word. He used the same words during study and test phases of the experiment; however, at the test, half of the words were presented in the same voice as during study and half were presented in a different voice. Goldinger found that identification of words in noise was most accurate when the voice was the same across study and test. Perceptual identification of words in noise revealed an advantage at all speaker levels for "same voice" trials, while changes in the speaker of the word across repetitions incurred costs. Further, the stimulus-specific details of a spoken word affected performance even after a week, demonstrating talker variability effects even across very long delays.[2]

Goldinger's findings suggest two possible conclusions. The first is that words to be identified are normalized in the conventionally understood way so that access to corresponding abstract lexical entries is possible. In addition, a representation of the spoken word is retained in an auditory or phonological PRS (Church & Schacter, 1994; Schacter & Church, 1992), or in episodic memory (Feustal, et al., 1983). A second possibility is that the lexicon itself is a form of exemplar memory composed of detailed traces of past experiences in which spoken words have been heard (Goldinger 1992; cf., Hintzman, 1986). Every encounter with a word results in the creation of a new memory trace. In this kind of a theory, the lexicon and episodic memory are not distinct memory systems. Individual exemplars serve as the basis for recognition of specific word tokens, and prototype knowledge is derived by retrieving multiple traces.

[2] These results are different from those reported by Schacter and Church (1992). They found no evidence for implicit voice effects when first occurrences of words were presented in the clear and second occurrences were presented in white noise. More recently, however, Church and Schacter (1994) did find reduced repetition effects when a speaker's voice, intonation or fundamental frequency changed between study items presented in the clear and test items that were low-pass filtered.

We do not know of any evidence that convincingly adjudicates between these views, but one consideration may favor the second. Accepting the first view (that two kinds of memory are accessed in word identification, one abstract and one instance-specific) implies that normalization takes place in two different ways, one for each memory system. To subserve lexical access, carrier information, such as voice, is stripped from phonological information about the word. In episodic memory, in order for a word in one voice to influence recognition or identification of a word in another, information for the consonants and vowels of a word must be *distinguished* from information for the voice. However, carrier and phonological information are not stripped away during encoding. The alternative conceptualization, that the lexicon itself is an exemplar or episodic memory system, requires just one kind of normalization, that of distinguishing information about consonants and vowels from information about the voice without eliminating voice information from memories for a word.

The purpose of our research is to explore further the nature of the memory traces for words spoken in different voices. In particular, we ask whether they are episodic memories in the commonsensical meaning of the term. That is, are the traces that support word recognition and identification memories for prior events or episodes in a perceiver's experience in which a word was spoken? If so, we should expect not only voice information about such an event to be preserved but also other event information.

## Experiment 1

The principal goal of Experiment 1 was to determine whether other aspects of an event in which a spoken word occurs, in addition to voice information, are also retained in long-term memory. Our question, specifically, was: Will optically specified information about a speech event be preserved along with information about a word and the voice that produced it? Accord-

ingly, will subjects better identify words as old when both the appearance and voice of the speaker uttering a word are preserved across presentations of a word?

*Method*

*Subjects.* Seventeen students enrolled in introductory psychology courses at the University of Connecticut volunteered to participate in the experiment in exchange for course credit. All subjects were native speakers of English with normal hearing and normal or corrected vision.

*Stimulus materials.* Twenty different speakers (10 males and 10 females) were videotaped uttering 280 monosyllabic words twice each. Following Palmeri et al. (1993) most of the words were selected from the Modified Rhyme Test (House, Williams, Hecker, & Kryter, 1965). Thirty additional monosyllables were selected from Kučera and Francis (1967).

Talkers were videotaped in a sound proof booth. Each word was presented in isolation at a fixed citation rate of 2.5 s on the CRT screen of a Macintosh computer. We adopted this procedure of presenting words singly and slowly on a computer terminal after trying several others. The procedure elicited speech that was fairly uniform both across speakers and within speakers across time. Further, although the procedure did not entirely eliminate intonation differences within and across subjects, largely it did so.[3] It also eliminated any list-final intonational falls in fundamental frequency, because the final list item was not detectable as such (i.e., it was distinguished from nonfinal items only in not being followed by another word).

Speakers produced the words in two blocks of 280 words, with one token of each word produced in each block. Sixteen of the 20 speakers (eight males and eight females) wore a hat or a scarf or both while producing one of the tokens of each word.

We then randomly selected (without replacement) 12 experimental tokens and two control fillers produced by each speaker (three from one of the speakers) to be used on the final stimulus tape. We created a test order in which experimental tokens were repeated at lags of 2, 4, 8, 16, 32, or 64 intervening items. Twenty experimental items were repeated at each lag.

Words were digitized at a sampling rate of 10.4 kHz. A second generation videotape was created by recording the visible speaker for each successive trial of the memory test onto a videotape and dubbing the voice tokens onto these. This was accomplished by allowing the onset of a speaker's acoustic signal on the original videotape to trigger a voice key, which signaled the computer to output a word onto the second generation videotape. The token was either the talker's own production of the visible word (the production that occurred originally with the face) or another talker's production. In this way, we were able to cross the independent variables voice (whether the voice was the same or different across presentations of a word) and face (whether the visible speaker was the same or different across presentations of a word). All trials were dubbed, even those in which the voice and face matched, and a given acoustic word was always dubbed onto a face of the same gender articulating that same word. The dubbing was quite compelling. In a task in which 29 subjects were asked to place check marks next to trials in which they noticed any asynchrony between the facial movements and the spoken word (see Experiment 2b be-

[3] Speakers used either a list intonation (final pitch flat or rising) or a declarative contour (final pitch falling). Both experimenters independently judged the intonation on each trial of our final test tape. We were in agreement that all ten male talkers used a declarative intonational contour consistently. Although female talkers used one or the other contour consistently, just half in judgments of the second author, or seven of the ten in judgments of the first author, used a declarative contour consistently. Across the two judges, the proportion of declarative contours ascribed to each of the talkers correlated significantly ($r = .827$).

low), subjects placed check marks on less than 4% of the 280 trials.

The final stimulus tape presented each visible speaker for approximately 500 ms before he or she began to speak. After the word was uttered, the visible speaker remained on the screen for another 500 milliseconds (approximately), followed by a black screen between trials. A new spoken word trial occurred three seconds after the offset of the visible speaker (a 3-s interstimulus interval (ISI)) except for trials that corresponded to the end of a column on the response sheet. For these trials, the ISI was 6 s. This allowed subjects to know when they should have reached the bottom of a column and hence to recover from errors in which they had missed a response. Overall intelligibility of the words (that is, percent correct identification of each word by ten introductory psychology students) was 84% (the range across listeners was 73–94%).

Three experimental factors were crossed as within-subject variables: voice, face and lag. Half of the repeated words were spoken in the same voice and the remaining half were spoken in a different voice on the second occasion. These trials were crossed with three face conditions. Specifically, the second presentation of a word was spoken by the same visible person, the "same+" person (the same individual had a hat and/or scarf during one trial), or a different person. Finally, the lag was also varied, such that repeated words were separated by 2, 4, 8, 16, 32, or 64 items.

One experimental token was inadvertently omitted from the tape, resulting in 119 first presentations of experimental trials (rather than 120). The omission affected the lags of 8 experimental tokens: Seven tokens were separated by 63 words rather than 64, and one token's lag changed from 32 to 31. On the final videotape, 119 words were first presentations to which the correct response was "new;" 119 words were repetitions of those words to which correct response was "old," and the remaining 42 words were new fillers. It was necessary to include fillers in order to achieve the intended lags between the first and second presentations of a word.

*Procedure.* A continuous recognition memory task based on word identity was administered to subjects. Subjects were tested in groups of three or fewer in a quiet room. They were told that they would be presented with 20 different speakers saying 280 words. After watching each word being spoken, subjects were instructed to indicate whether the word had occurred previously on the videotape by circling new or old on an answer sheet. They were told that a word was to be identified as an old token even if it was repeated by a different person. They were also told that the lag between first and second presentations of a word would vary. The session lasted 30 min.

### *Results*

Recognition accuracy was measured as overall proportion correct (proportion old responses) on the second presentations of words. In Fig. 1, proportion correct old responses is presented as a function of the lag between the first and second occurrences of a word. The data are presented separately for same and different voice trials, and, in different panels of the figure, for the three face conditions. Because there are so few items per lag, we grouped our six lags into three categories: short (lags 2 and 4), medium (lags 8 and 16) and long (lags 32 and 64). Figure 1 shows that we replicated the important findings of Palmeri et al. in that accuracy identifying a word as "old" was greater when voice was preserved between the first and second occurrences of a word than when it was not. The overall advantage of voice preservation was 5%. A repeated measures analysis of variance (ANOVA) was performed with factors: voice (same voice, different voice), face (same person, same+ person, different person) and lag. The main effect of voice [$F(1, 16) = 15.46$, $MS_e = .01$] was highly significant. In addition, there

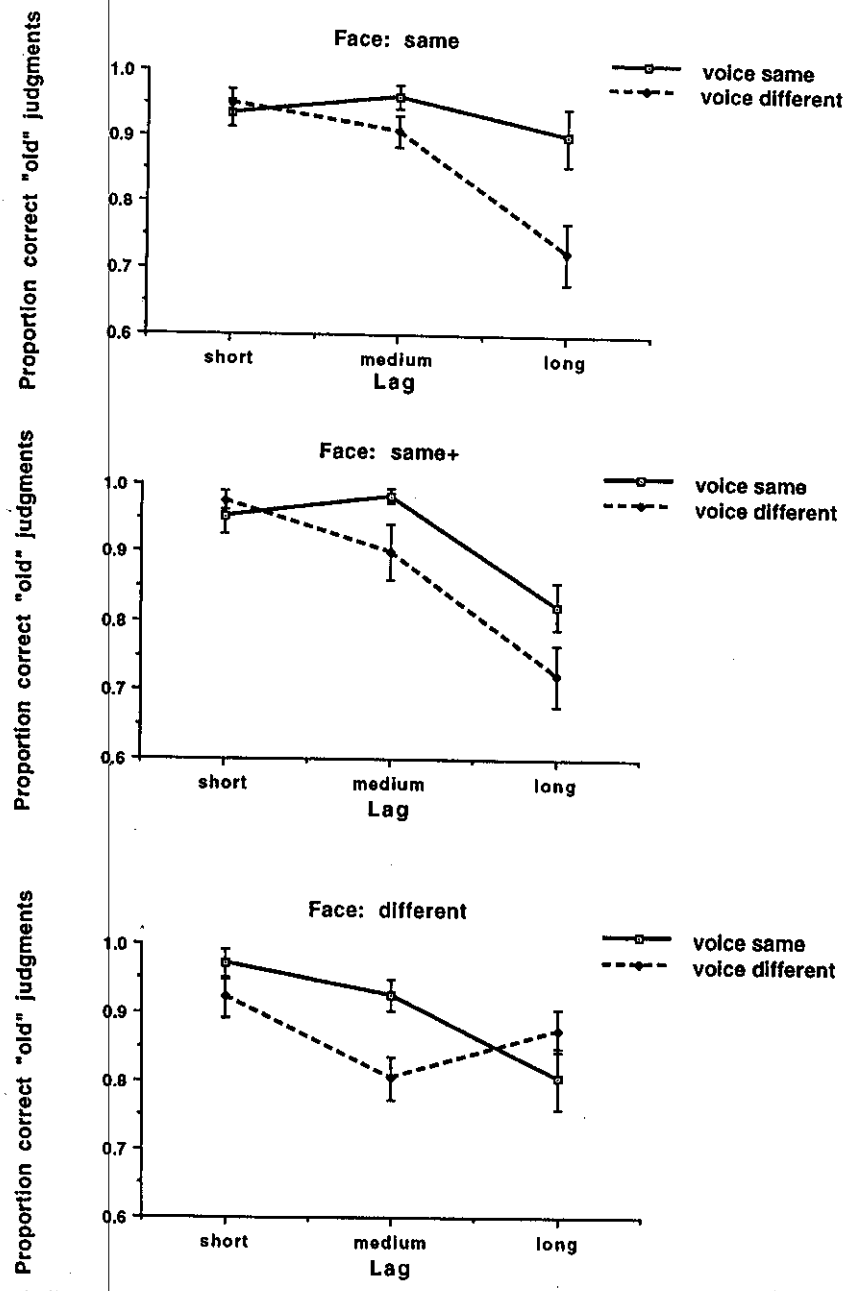


FIG. 1. Recognition accuracy is presented as a function of the lag between the first and second occurrences of a word; data are presented separately for same and different voice trials. The first panel displays item recognition for same face trials as a function of voice and lag conditions. The second panel displays item recognition for same+ face trials. The lower panel displays item recognition for different face trials. (Each mean is based on 17 data points.)

was a significant effect of lag [$F(2, 32) = 22.73$ $MS_e = .02$], reflecting a consistent decrease in recognition accuracy as lag increased.

In contrast to our significant effect of voice and lag, we found no effect of the face manipulation on recognition accuracy ($F < 1$). That is, subjects were only slightly more

accurate identifying a word as old when the same face uttered the word on new and old presentations of the word, relative to trials on which a different face produced the word on one occasion (a 1.4% difference).

There were two significant two-way interactions, between voice and lag and between face and lag [$F(2, 32) = 3.52$, $MS_e = .01$ and $F(4, 64) = 5.35$, $MS_e = .01$, respectively], and a significant three-way interaction of voice, face and lag [$F(4, 64) = 6.31$, $MS_e = .01$]. The three-way interaction does not lend itself to a straightforward interpretation. However, as the figure shows there was an increase in the magnitude of the voice condition over lags on "same face" (and to a lesser extent, on the "same+") trials which probably reflects movement of performance off ceiling at the longer lags. In contrast, there was a nonmonotonic change in the voice effect across lags on different-face trials, such that the voice effect increased numerically from short to medium lags, but was absent, even reversed, at the longest lag. This finding of implicit voice effects on word recognition at long lags, but only in conditions in which the face is the same, can be interpreted as an implicit face effect that is only apparent at long lags. However, because we have just one test order, we cannot rule out the possibility that the pattern is caused by trial-to-trial idiosyncrasies in our test tape.

To ascertain whether trial-by-trial idiosyncrasies were responsible for creating any of the significant factors in the subjects analysis, we conducted an items analysis with the same factors as in the analysis by subjects. Indeed, in this and all other experiments, the items analysis yields only significant main effects of voice and lag, and no significant interactions. Accordingly, we cannot know that, with another set of items (or even the present items rearranged), the three-way interaction would emerge as it did in the subjects analysis.

*Discussion*

In Experiment 1, we replicated the findings of Craik and Kirsner (1974) and of Palmeri et al (1993) by showing that subjects were more accurate at recognizing old words as such when the same voice, rather than a different voice, produced the word on the first and second occasions. Our result reinforces the idea that voice information is retained in long-term memory with information for a word's identity. Furthermore, the data also indicate that the memory that subserves explicit word recognition includes information about voice. However, we found no analogous effects of face specificity, in that preserving the face across study and test did not influence word recognition. The only possible hint of an effect of face that we obtained was the finding in the three-way interaction that the effect of the voice on word-recognition judgments had disappeared at the longest lags in the different face condition but was still present in the same face condition.

One reason for our findings may be that optical information was not encoded (or, if our three-way interaction is meaningful, was only weakly encoded) during the task. Possibly, optical information is never encoded with the memory for a speech event. Alternatively, our experimental design may have fostered inattention to the face in two ways. First, as discussed above, the dubbing of different voices onto different faces may have disrupted the use of visual information, causing subjects to rely predominantly on the acoustic signal and to ignore the visual information. Although we cannot rule this factor out entirely, two considerations suggest to us that the dubbing was not the reason that we did not obtain a face effect. First, our dubbing was virtually undetectable in that very few of our trials were deemed asynchronous by subjects (see Experiment 2b below). A second consideration is the McGurk effect (e.g., McGurk & MacDonald, 1976), in which an acoustic speech signal is dubbed onto a dif-

ferent visibly mouthed syllable or word, resulting in the perception of a single, fused sound. Importantly, awareness of the bimodal discrepancy does not cause subjects to ignore the visual information (e.g., Liberman, 1982). A second, more plausible, source of inattention to the optical information could simply be the nature of the task. That is, our continuous recognition task could have been performed successfully with or without attending to the visible speaker. Therefore, before concluding that optical information about a speech event is generally not stored in long-term memory along with voice and lexical information, we redesigned the task so that attending to the video display was mandatory.

## Experiments 2a and 2b

In Experiments 2a and 2b, we selected two different secondary tasks for subjects to perform, both designed to require attention to the visible speaker. In one condition, after making their new/old judgment, subjects judged whether a speaker's hair was long, short, or average for his or her gender. The second condition focused subjects' attention on the visible articulation, so that they attended to the face at the moment the word was being uttered. In this case, subjects judged the synchrony of the mouth movements relative to the acoustic signal.

### *Method*

*Subjects.* Fifteen students enrolled in introductory psychology courses at the University of Connecticut served as subjects in Experiment 2a, and 29 subjects from the University of Alaska volunteered to be subjects in Experiment 2b, each group in exchange for course credit. All subjects were native speakers of English with normal hearing and normal or corrected vision.

*Stimulus materials.* The stimulus materials for Experiments 2a and 2b were those used in Experiment 1.

*Procedure.* The general procedure for Experiments 2a and 2b was similar to that of Experiment 1. The only change was that subjects were required to make an additional judgment after each word recognition judgment.

In Experiment 2a, subjects were asked first to judge whether the word they heard was new or old and next to assign a number (1, 2, or 3, respectively) reflecting whether, for the person's gender, his/her hair length was shorter than average, average, or longer than average. It happened that our videotaped speakers had hair lengths that ranged from very short, to very long for each gender. Accordingly, the hair judgments were feasible.

In Experiment 2b, subjects were asked to judge first whether the word they heard was new or old. Next, they were instructed to put a check next to those trials in which the timing of the acoustic information was out of synchrony with the visible mouth movements. This task, in contrast to that of Experiment 2a, required subjects to attend to the face at the same time that they attended to the audible spoken word.

### *Results*

We first asked whether our new instructions changed performance with respect to the effects of voice and lag.

Table 1 lists the mean proportion correct old" responses for each experiment, across both levels of the voice factor. The table makes clear two points. The first is that the overall benefit for same voice repetitions in Experiments 2a and 2b is similar to that in Experiment 1. Second, our two instruction sets had no overall effect on recognition performance ($F < 1$). Further, the instruction set factor did not interact significantly with any other factor in the design. Accord-

TABLE 1
Mean Proportion Correct "Old" Responses as a Function of the Voice Factor

| | Expt. 1 | Expt. 2a | Expt. 2b |
|---|---|---|---|
| Voice same | .92 | .91 | .92 |
| Voice different | .87 | .86 | .84 |

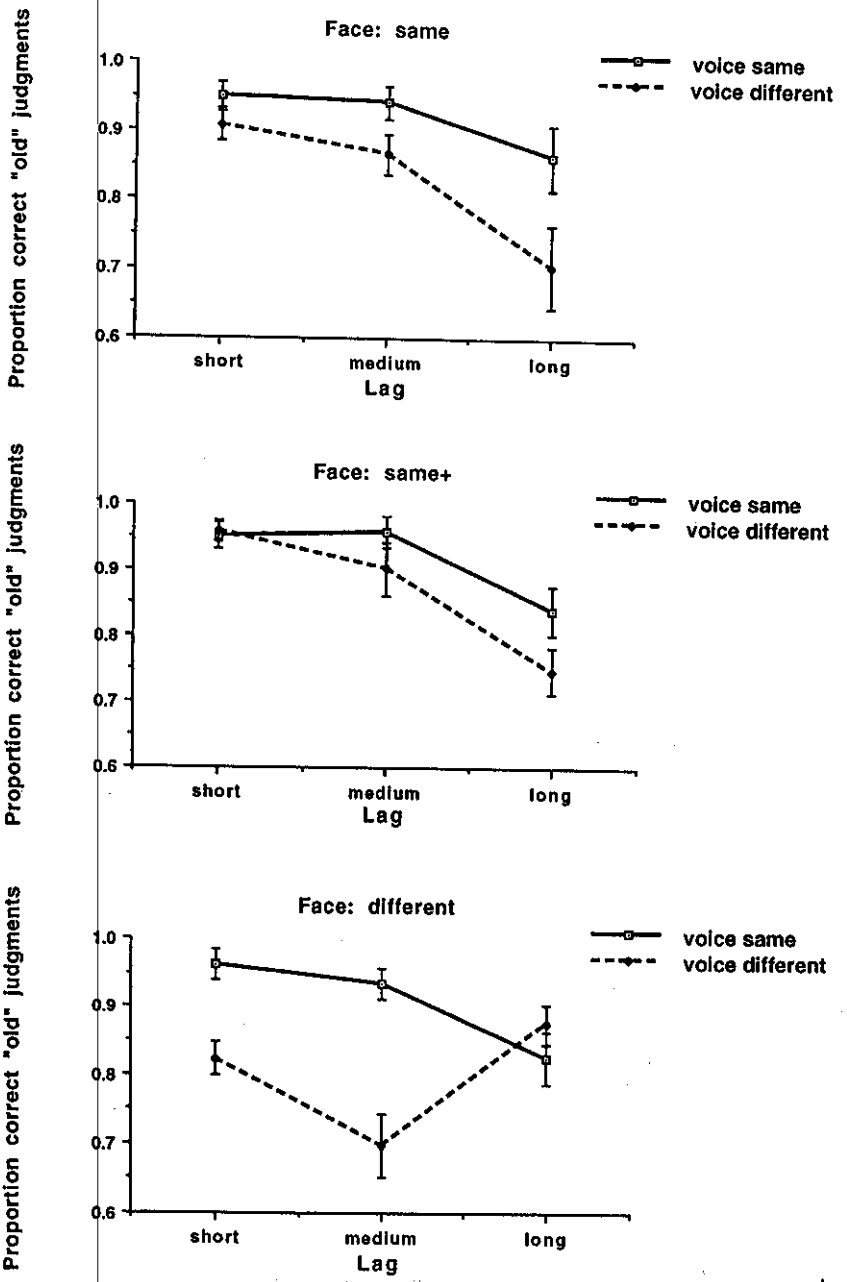


FIG. 2. The combined word recognition accuracy for Experiments 2a and 2b is presented as a function of the lag between the first and second occurrences of a word; data are presented separately for same and different voice trials. The first panel displays item recognition for same face trials as a function of voice and lag conditions. The second panel displays item recognition for same+ face trials. The lower panel displays item recognition for different face trials. (Each mean is based on 22 data points.)

ingly, we present the data combined from the two experiments.

Figure 2 presents the data separately for same and different voice trials for the three face conditions. As the figure illustrates, there were significant effects of voice [$F(1, 43) = 82.47$, $MS_e = .01$] and lag [$F(2, 86) = 54.24$, $MS_e = .02$]. Recognition was more

accurate when the voice was preserved and at shorter lags.

Although there is a numerical trend for word recognition to improve when the visual information is most similar across presentations (averaging 1.4% comparing recognition performance on same face and different face trials), the effect did not approach significance ($F < 1$). There was a significant interaction between face and lag [$F(4, 172) = 11.48$, $MS_e = .01$], and a three-way interaction between voice, face and lag, [$F(4, 172) = 15.50$, $MS_e = .01$]. As before, the voice effect grew over lags in the same face condition, but it disappeared at the longest lags in the different voice condition. In addition, there is a paradoxical outcome that the voice effect is numerically largest of all at the short and medium lags of the *different* face condition.

Given that we again found no evidence for an implicit face effect in these experiments (or little evidence for it if our three-way interaction is interpreted as showing a loss of the voice effect at long lags in the different-face condition), it is important to ask whether subjects in fact took their secondary task seriously. On the hair judgments of Experiment 2a, we ran a repeated measures analysis of variance on the judgments using visible speaker as a factor. We found a highly significant effect of visible speaker [$F(19, 266) = 36.224$, $MS_e = .19$] such that visible speakers with long hair for their gender were judged to have longer hair than those with shorter hair. The five speakers judged by the experimenters to have the longest hair for their gender were given average ratings near the maximum of 3 ($M = 2.74$). In contrast, the seven speakers judged by the experimenters to have the shortest hair were given average ratings near the minimum of 1 ($M = 1.21$). On the dubbing judgments of Experiment 2b, as noted earlier, subjects checked fewer than 4% of the trials as asynchronous. On the two trials that the experimenters judged the least and second least successfully dubbed, 19 and 10 of the 29 subjects, respectively, placed checkmarks next to those trial numbers. Thus, we have clear evidence, particularly in Experiment 2a, that subjects were attending to the visible speakers as instructed.

### *Discussion*

A conclusion that is increasingly tempting to draw from the findings of Experiments 1 and 2 is that, whereas acoustic information other than that about the component consonants and vowels is preserved in memory for a word, optical information—whether about speech gestures or merely speaker appearance—is not. This conclusion assumes, however, that the absence of implicit effects of face preservation on new/old judgments of words means that face information is not preserved with information for the spoken word. The next experiment tests that assumption.

In Experiment 3, we attempt to force subjects to preserve information about the visible speaker with information for each spoken word. Let us assume that we succeed. If the assumption is correct that implicit effects of face preservation will occur when visible speaker information is preserved in memory, then we should observe facilitation in recognition of a spoken word as old if the visible speaker is the same across first and second presentations of a word. However, this assumption may not be correct. We may find that, although subjects can retain information about the visible speaker of the word, there are still no implicit improvements in word recognition. If that occurs, we will know that preservation of event information in memory does not guarantee implicit improvement in word recognition when event information is retained.

## Experiment 3

We encouraged subjects to preserve information for the visible speakers in memories for the spoken words by having them

make judgments about face-word pairings on "old" trials. That is, when they judged a word to be old, they made a second judgment whether the face saying the word on the second occurrence was the same or different as the face saying the word on it's first occurrence. These second judgments could only be made with better than chance accuracy if information for visible speakers is preserved with information about the words they spoke.

*Method*

*Subjects.* Seventeen students enrolled in introductory psychology courses at the University of Connecticut volunteered to participate in the experiment in exchange for course credit. All subjects were native speakers of English with normal hearing and normal or corrected vision.

*Stimulus materials.* The stimulus materials for Experiment 3 were those used in Experiments 1 and 2.

*Procedure.* The general procedure for Experiment 3 was similar to that of the previous experiments. However, in Experiment 3, subjects made a three-way judgment. They responded "new" if the word was new in the experiment; they responded "old–same" if the word was old and the visible person was the same on both presentations; otherwise, they responded "old–different." We instructed subjects to respond old–same on "same+" pairings where the speaker was the same person, but he or she looked a little different on the new and old trials.

*Results*

*Item recognition.* For the analysis of item recognition, judgments about the visible speaker (old–same and old–different) were pooled to created a single "old" category. This allowed comparison of word recognition performance to be made across experiments. In this experiment, subjects were overall correct on 92% of the trials when the voice was preserved and 84% correct when the voice was not. These values are identical to those for Experiment 2b, shown in Table 1. Thus, the additional explicit visible speaker judgment did not change the voice effect or the overall accuracy in the recognition task.

In Fig. 3, recognition accuracy is presented as a function of the lag between the first and second occurrences of a word. The data are presented separately for same and different voice trials, and, in different panels of the figure, for the three face conditions. As in the previous experiments, a repeated-measures analysis of variance revealed a significant effect of voice: preservation of voice across trials improved word recognition accuracy [$F(1, 16) = 24.85$, $MS_e = .02$]. Likewise, there was a significant effect of lag [$F(2, 32) = 23.70$, $MS_e = .02$] whereby recognition performance decreased as lag increased. There was also a significant interaction between voice and lag [$F(2, 32) = 3.55$, $MS_e = .01$].

As in our previous experiments, we found a nonsignificant main effect of face, with recognition judgments on same face trials more accurate than those on different face trials by just 1.0% [$F(2, 32) = 1.53$, $MS_e = .02$]. The interactions of face and lag [$F(4, 64) = 5.20$, $MS_e = .01$] and of face, voice, and lag were significant [$F(4, 64) = 7.40$, $MS_e = .01$]. As in our previous experiments, the voice effect grew over lags in the same face condition, but it peaked at the middle lag condition in the different-face condition.

*Face judgments.* Given that we once again found no consistent implicit effect of preserving the face on new/old word recognition judgments, it is of interest to ask whether subjects retained information about the face and word pairings. To assess their explicit memory for face and word pairings, we conducted analyses on the old–same and old–different responses—that is, on judgments of whether the visible speaker was the same or different on both presentations of a repeated word. This was accomplished by comparing the proportion of correct visible speaker recognitions on

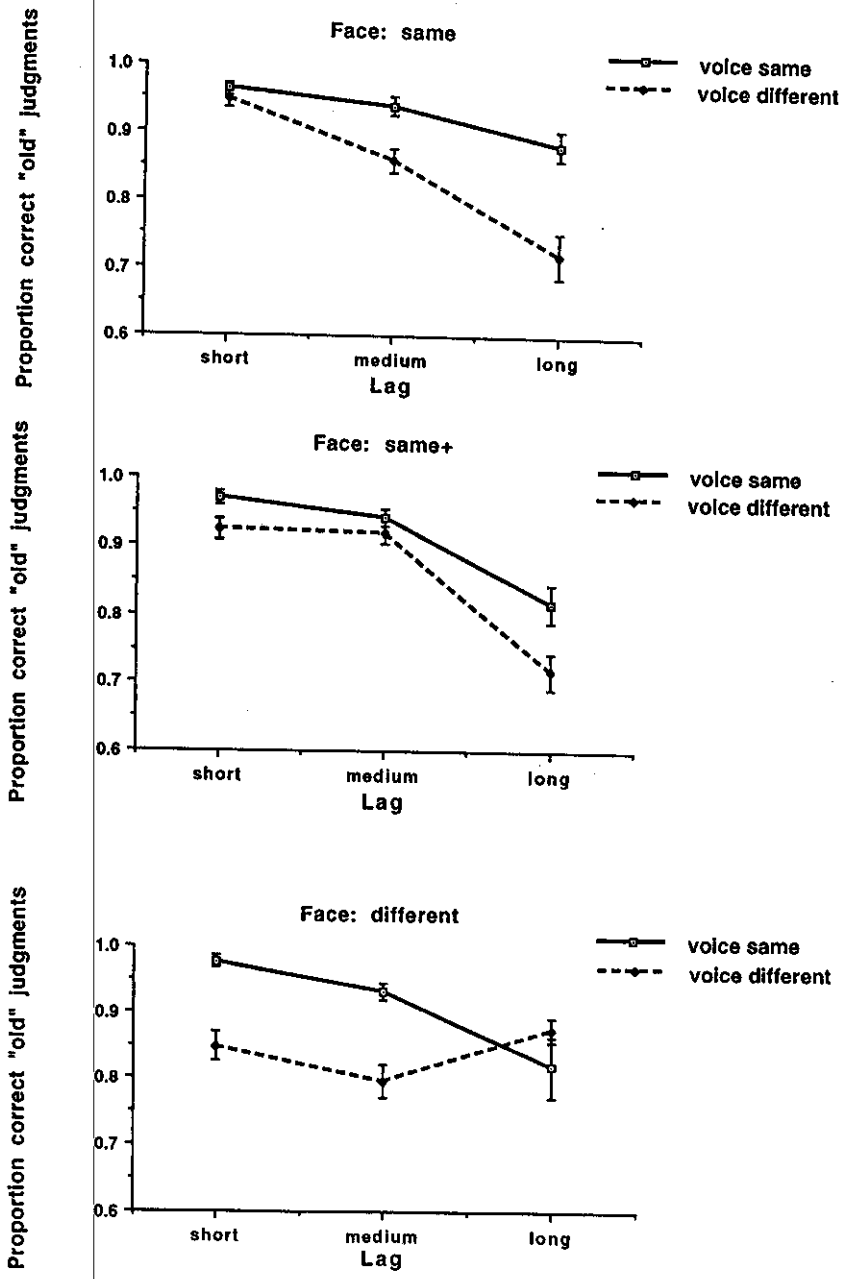


FIG. 3. Recognition accuracy is presented as a function of the lag between the first and second occurrences of a word; data are presented separately for same and different voice trials. The first panel displays item recognition for same face trials as a function of voice and lag conditions. The second panel displays item recognition for same+ face trials. The lower panel displays item recognition for different face trials. (Each mean is based on 17 data points.)

trials on which the word had been correctly identified as old with the total number of correct speaker recognitions that would be expected by chance. Chance recognition was defined for each subject as 50% of the total number of his or her correct old judgments. These findings are presented in Fig. 4. Subjects were able to make the same/

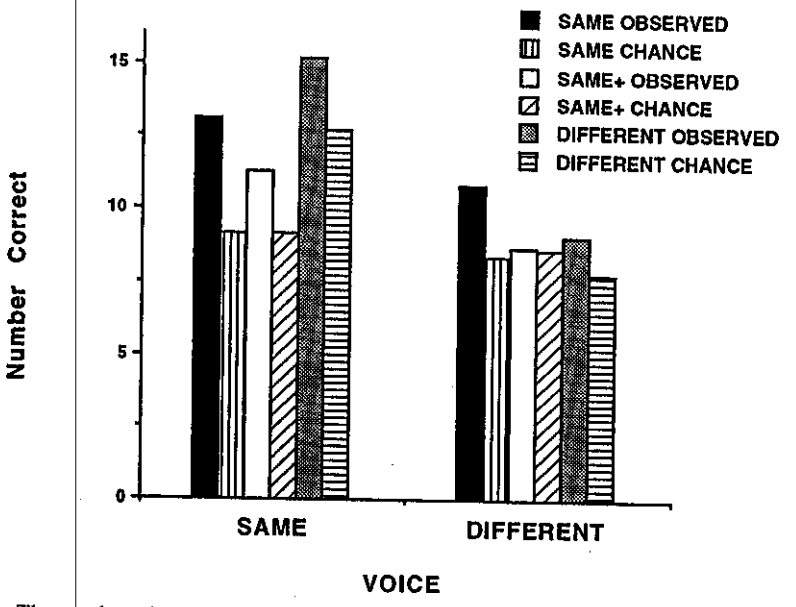


FIG. 4. The number of correct observed "old–same" or "old–different" explicit face judgments and their corresponding chance number of judgments are plotted as a function of voice and face conditions, collapsed across levels of lag. (Each mean is based on 17 data points.)

different-speaker judgments better than chance [$F(1, 16) = 104.63$, $MS_e = 5.10$]. The interaction between the face condition and observed/chance performance was significant [$F(2, 32) = 21.45$, $MS_e = 3.73$], however. Although subjects were significantly above chance in all three face conditions, their departure from chance was smallest in the same+ condition (where the correct judgment was same to faces that looked somewhat different owing to the presence in just one occurrence of a hat and/or scarf) and it was largest in the different condition.

A different interesting feature of the outcome is that we obtained a significant implicit influence of voice on the same/different visible speaker judgments. That is, bars on the left half of Fig. 4, representing observed same voice conditions, exceed bars reflecting chance performance to a greater extent than do corresponding bars on the right side representing observed different voice conditions. This significant interaction between voice and the observed/chance factor [$F(1, 16) = 14.27$, $MS_e = 2.20$] means that subjects are significantly more accurate both identifying the same visible speaker as "same" and a different speaker as "different" when voice is preserved across presentations of a word. The effect of voice preservation, then, is not to bias a response of "same" when voice is preserved, but rather to improve accuracy on both "same" and "different" judgments. The three way interaction of voice, face and observed/chance was not significant.

*Discussion*

In Experiment 3, we have evidence that, under instructions to do so, subjects can retain visual information about an event in which a spoken word occurs along with information about the word itself. However, even under these conditions, we did not find a significant improvement in recognition of spoken words as old when face information was preserved across trials.

The lack of implicit face effects in this experiment, as in Experiments 1 and 2 (excepting the loss of the voice effect at long

lags in the different face condition), leads us to ask why we do get an improvement in word recognition when voice is preserved, but do not when the face is preserved. Perhaps this pattern relates to the finding that, under certain conditions, spoken words and voices are processed integrally, whereas spoken words and faces may not be. For example, Mullenix and Pisoni (1990) found that voice changes interfered with subjects' attempts to attend selectively to phonetic information in a speeded classification task, suggesting that words and voices are processed in a mutually dependent fashion. Although it has not, to our knowledge, been tested, possibly under conditions similar to our own, words and faces are not processed integrally.

Another interpretation of the findings in Experiment 3 is that, whereas under instructions to retain face information in a speech event subjects can do so, possibly their memory for faces is not as robust as their memory for voices in the context of a linguistic task. Perhaps such a difference in the strength of memories for faces and voices underlies differences in their implicit effects on word recognition judgments. Experiment 4 was designed to permit a comparison of explicit memory for the voice and the face.

## Experiment 4

Experiment 4 is a replication and extension of Experiment 2 of Palmeri, et al. (1993). In that experiment, subjects were asked to identify words as new or old and, for old judgments, to make a secondary judgment as to whether the voice producing the word was the same or different on both repetitions. Palmeri, et al. (1993) found that subjects were able to make the same/different judgments about the voice with better than chance accuracy even at the longest lags. Our experiment has a similar design, but includes the face variable.

### *Method*

*Subjects.* Twenty-two students enrolled in introductory psychology courses at the University of Connecticut volunteered to participate in the experiment in exchange for course credit. All subjects were native speakers of English with normal hearing and normal or corrected vision.

*Stimulus materials.* The stimulus materials for Experiment 4 were those used in the previous three experiments.

*Procedure.* In Experiment 4, as in Experiment 3, subjects made a three-way judgment. This time they responded "new" if they judged that the word was being presented for the first time on the tape. They responded old–same if they judged that word as both old and repeated in the same *voice;* otherwise, they responded old–different.

### *Results*

*Item recognition.* For the analysis of word recognition, voice judgments ("old-same" and "old-different") were pooled to created a single "old" category. This allowed comparisons of word recognition performance to be made across experiments (see Table 1). In this experiment, subjects were overall correct on 91% of the trials when the voice was preserved and 86% correct when the voice was not. These values are comparable to those for the previous experiments. Thus, the additional explicit voice judgment did not change the magnitude or direction of the voice effect.

In Fig. 5, recognition accuracy is presented as a function of the lag between the first and second occurrences of a word. The data are presented separately for same and different voice trials, and, in different panels of the figure, for the three face conditions. As in the previous experiments, a repeated-measures analysis of variance revealed a significant effect of voice: preservation of voice across trials improved word recognition accuracy [$F(1, 21) = 23.60$, $MS_e = .01$]. Likewise, there was a highly significant effect of lag [$F(2, 42) = 54.87$, $MS_e = .01$] whereby recognition performance decreased as lag increased.

As in our earlier experiments, there was

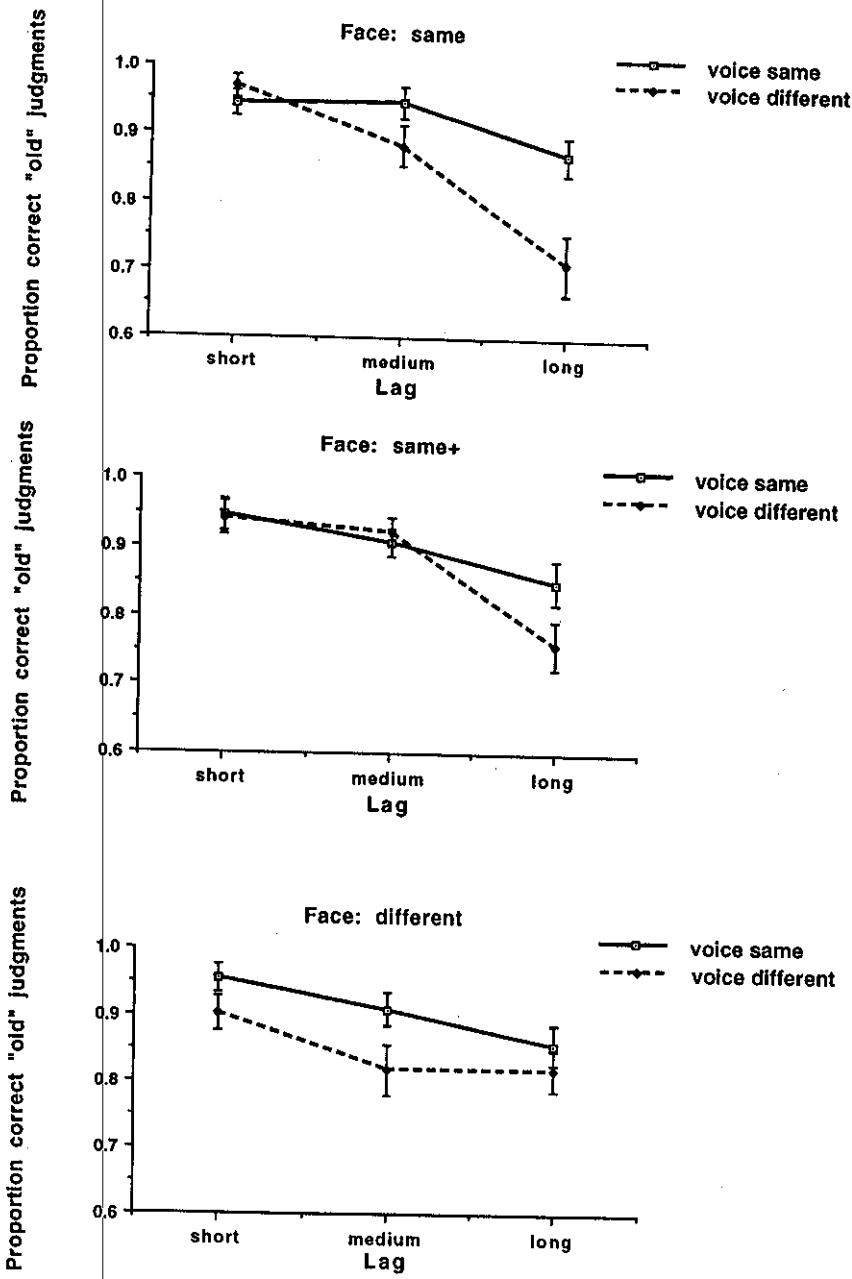


FIG. 5. Recognition accuracy is presented as a function of the lag between the first and second occurrences of a word; data are presented separately for same and different voice trials. The first panel displays item recognition for same trials as a function of voice and lag conditions. The second panel displays item recognition for same + face trials. The lower panel displays item recognition for different face trials. (Each mean is based on 22 data points.)

a nonsignificant effect of face preservation on word recognition ($F < 1$) with word recognition performance on "same" face trials exceeding that on "different" face trials by 1.1%. There was a significant interaction between lag and voice [$F(2, 42) = 4.10$, $MS_e = .02$] lag and face [$F(4, 84) = 2.51$, $MS_e = .01$] and a three way interaction of

voice, face, and lag [$F(4, 84) = 3.60$, $MS_e = .01$]. Consistent with our previous experiments, there is a trend for the voice effect to grow over lags in the same face condition, and disappear at the longest lags in the different voice condition.

*Analysis of implicit face effects across experiments.* Although in none of the experiments has the implicit face effect approached significance, the overall direction of the effect is uniformly as predicted. On average the difference between the same face and different face conditions was 1.4, 1.2, 1.5, 1.0, and 1.1% for Experiments 1, 2a, 2b, 3, and 4, respectively. A paired, two tailed t-test comparing each subject's face effect against zero yields a significant effect when data from all experiments are collapsed [$t(99) = 2.284$, $RMS_e = .061$]. This is a clear indication that visual face information is preserved with information about spoken words. Experiment 4 provides a second indication as well.

*Voice judgments.* Analyses were conducted to assess subjects' explicit voice judgments on the "old" responses. This was accomplished by comparing the number of correct voice recognitions with the number of correct voice recognitions that would be expected by chance. Chance recognition was defined for each subject as 50% of the total number of his or her correct "old" judgments. Performance was significantly above chance identifying the voice as the same or different [$F(1, 21) = 24.19$, $MS_e = 4.66$]. There was a significant interaction between the voice conditions and the chance factor [$F(1, 21) = 5.41$, $MS_e = 7.94$]. Interpretability of this interaction is affected, however, by a highly significant three-way interaction between voice, face and observed/chance [$F(2, 42) = 32.90$, $MS_e = 5.85$]. This is shown in Fig. 6.

As the figure shows, among the six combinations of the variables face and voice, performance in three conditions is numerically above chance whereas performance in the other conditions is numerically below chance. This is in contrast to findings in Experiment 3 (Fig. 4) in which performance making face judgments is numerically above chance in all six conditions.

The patterning of above- and below-chance performances in Experiment 4 is

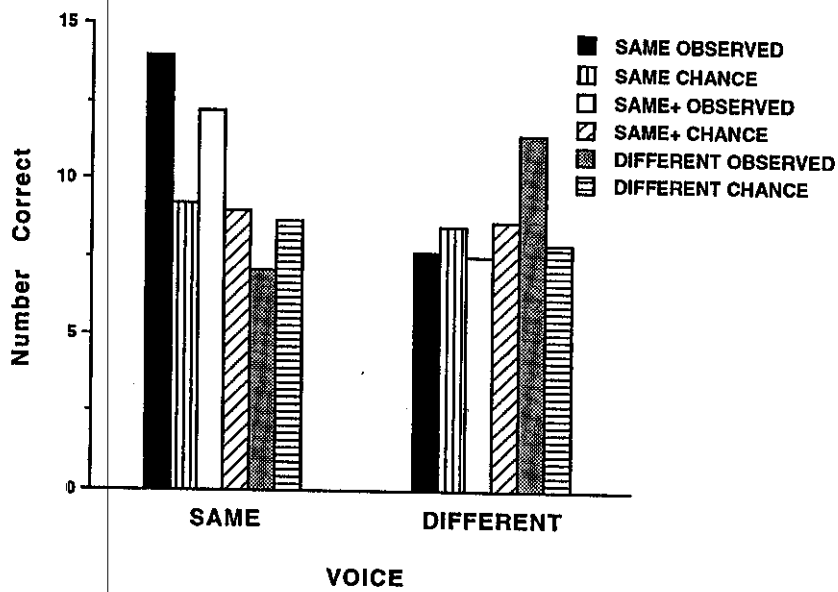


FIG. 6. The number of correct observed "old–same" or "old–different" explicit voice judgments and their corresponding chance number of judgments are plotted as a function of voice and face conditions, collapsed across levels of lag. (Each mean is based on 22 data points.)

easily described. Subjects performed above chance in judging that the voice was the same across presentations of a word when the face was the same, but they performed below chance when the face was not the same. Compatibly, subjects performed above chance judging that the voice was different across repetitions of a word when the face had not been preserved. They performed below chance on different voice trials where the face was preserved across presentations of a word. These findings are in sharp contrast to the analogous conditions of Experiment 3 in which we found improvement, due to preserving the voice, on judgments whether the visible speaker was the same or different across presentations of a word. In that experiment, preserving the voice improved both judgments that the visible speaker was the same across presentations of a word and judgments that the visible speaker was different. In the present experiment, implicit influences of memory for the face on voice judgments appear to be biasing effects. Subjects are biased to judge the voice the same across presentations of a word if the face is the same and to judge the voice different if the face is different. This improves accuracy when voice and face conditions match, but drives performance below chance when they do not. We will speculate on reasons for this outcome under Discussion and under General Discussion below.

*Discussion*

Our intention in Experiment 4 was to compare explicit memory for voice with explicit memory for the face obtained Experiment 3. We speculated that our consistent findings of implicit influences of voice preservation on word recognition, and of markedly weaker effects of the face, could be ascribed to a more robust memory for the voice than for the face. Our findings in Experiment 4 do not support that idea. Although the memory for the voice with which a word is spoken is retrieved with an accuracy better than chance, the accuracy level is lower, not higher, than for retrieval of the face associated with a word. The reasons why voice preservation (or the lack) strongly affects word recognition, but face preservation (or the lack) does so to a considerably lesser extent, must be sought elsewhere.

A new finding of the experiment is that face information was preserved along with a memory for a spoken word even when no special instructions were given to encourage subjects to attend to the face. Instructions to subjects regarding attention to the television screen were like the instructions given to subjects in Experiment 1. Subjects were asked to watch the television, but nothing about the task compelled attention to it. Here, as in Experiments 1–3, we see very weak evidence of implicit face effects on word recognition, but we do see a strong effect of the face that our measures in Experiments 1 and 2 could not have exposed. Information was retained about the visible speaker producing a word that exerted a marked biasing effect on explicit judgments of whether voice was preserved across presentations of a word. Because, as in Experiment 1, no special instructions were given to subjects compelling them to attend to the television screen, we are disposed to conclude that visual information about the faces producing words was preserved in the earlier experiments as well. Our measures were simply the wrong ones to reveal preservation of face information.

Even if this is the case, our findings in Experiment 4 also suggest that face information is preserved differently from voice information: Voices are preserved across repetitions of a word in such a way that word recognition judgments and judgments about the face producing a word on both repetitions are improved in accuracy. In contrast, preservation of the face has little effect on word recognition judgments. However, it does bias subjects to judge that voice was preserved whether it was or not. Compatibly, the failure to preserve the face across presentations of a word biases judg-

trast to these strong implicit effects of the voice on recognition memory performance, explicit memory for voice–word pairings is rather poor. As for the face, we find very weak improvement in word recognition when a face is preserved across new and old presentations of a word, and we find biasing effects of face preservation on recognition memory for voice–word pairings. In contrast to the weak (or, in the case of the voice–word pairings, the malign) implicit effects of the face, explicit memory for the face is good. We do not know how to account for this particular patterning in the data. We presume that it has something to do with the attentional focus on word information that our recognition memory task fosters. Further research is necessary to make sense of this provocative patterning of results.

An alternative to our conclusion that normalization is not a process of stripping nonphonological information from phonological information might be developed from the perspective of Schacter's PRS ("presemantic representational systems" e.g., Schacter, 1992). Stimulus input leaves a perceptual trace in any of several subsystems of the PRS. The subsystems are modality-specific, and there may be distinct auditory and phonological subsystems. If phonological and auditory subsystems are distinct, then the perceptual records have been normalized in the conventional sense. The reason why we see larger implicit voice than face effects on memory for spoken words, then, may be that the linkage between the two within-modality subsystems is closer than between the visual subsystem and the phonological subsystem. This account does not explain all of our findings but then neither does our own. The most puzzling thing about our findings for this account, however, must be why we see implicit effects at all in a recognition memory task.

Church and Schacter (1994) provide evidence for a double dissociation between variables affecting implicit and explicit memory tasks that they interpret as strong evidence that performance on the two tasks is served by different memory systems. Specifically, during study, words were presented in the clear; during test, they were low-pass filtered and the subjects' task was to identify them. Voice changes between study and test lowered performance identifying the low-pass filtered words, but voice changes had no effect on a subsequent recognition memory test. In contrast, a later experiment that manipulated levels of processing during study affected recognition memory but not identification of the filtered words. Ostensibly, performance on the perceptual identification task is mediated by PRS whereas recognition memory performance is based on an episodic memory. Although voice information should be represented in the episodic memory, Church and Schacter propose that "subjects tend to rely on conceptually driven processes when engaging in explicit retrieval" (p. 532). Accordingly, perceptual information, such as information for the voice, will not affect task performance. From this perspective, our findings, those of Palmeri et al. (1993), of Craik and Kirsner (1974), and the explicit-memory findings of Goldinger (1992), all of which showed effects of voice changes in explicit memory tasks are unexpected. Accounting for them would require explaining why those tasks and not the recognition memory task of Church and Schacter led subjects to rely more on data-driven than conceptually driven processing. It is not obvious to us why they should do so.

For us, the balance of the evidence still favors a view of normalization as distinguishing information about the phonological properties of speech events from information about nonphonological aspects, with no loss of nonlinguistic information, and a view of the lexicon as episodic. Given that nonlinguistic information is not stripped away in episodic memory, but, to be useful, the two kinds of information, linguistic and nonlinguistic, must certainly be

distinguished by perceivers, a view that conventional normalization occurs and that the lexicon is a memory of abstract types implies, as we noted in the introduction, that two normalization processes occur to subserve two kinds of memory systems. We find this implausible.

## References


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

Craik, F. I. M., & Kirsner, K. (1974). The effects of speaker's voice on word recognition. *Quarterly Journal of Experimental Psychology*, **26**, 274–284.

Feustel, T. C., Shiffrin, R. M., & Salassoo, A. (1983). Episodic and lexical contributions to the repetition effect in word identification. *Journal of Experimental Psychology: General*, **112**, 309–346.

Fowler, C. A. (1994a). Speech perception: Direct realist theory. *Encyclopedia of language and linguistics*. (Vol. 8, pp. 4199–4203). Oxford: Pergamon Press.

Fowler, C. A. (1994b). Invariants, specifiers, cues: An investigation of locus equations as information for place of articulation. *Perception and Psychophysics*, **55**, 597–610.

Gerstman, L. H. (1968). Classification of self-normalized vowels. *IEEE Transactions on Audio and Electoacoustics*, **au-16**, 630–640.

Goldinger, S. D. (1992). Words and voices: Implicit and explicit memory for spoken words. Unpublished Doctoral Dissertation, Indiana University.

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review*, **93**, 411–428.

House, A. S., Williams, C. E., Hecker, M. H. L., & Kryter, K. D. (1965). Articulation-testing methods: Consonantal differentiation with a closed-response set. *Journal of the Acoustical Society of America*, **37**, 158–166.

Jackson, A., & Morton, J. (1984). Facilitation of auditory word recognition. *Memory & Cognition*, **12**, 568–574.

Johnson, K. (1990). The role of perceived speaker identity in F0 normalization of vowels. *Journal of the Acoustical Society of America*, **88**, 642–654.

Joos, M. A. (1948). Acoustic Phonetics. *Language*, **24**(Suppl. 2), 1–136.

Kučera. F., & Francis, W. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown Univ. Press.

Ladefoged, P., & Broadbent, D. E. (1957). Information conveyed by vowels. *Journal of the Acoustical Society of America*, **29**, 98–104.

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

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

Mullennix, J. W., & Pisoni, D. B. (1990). Stimulis variability and processing dependencies in speech perception. *Perception & Psychophysics*, **47**, 379–390.

Nearey, T. M. (1989). Static, dynamic, and relational properties in vowel perception. *Journal of the Acoustical Society of America*, **85**, 2088–2113.

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 and Cognition*, **19**, 309–328.

Richardson-Klavehn, A., & Bjork, R. A. (1988). Measures of memory. *Annual Review of Psychology*, **39**, 475–543.

Schacter, D. L. (1992). Priming and multiple memory systems: Perceptual mechanisms of implicit memory. *Journal of Cognitive Neuroscience*, **4**, 244–256.

Schacter, D. L. & Church, B. A. (1992). Auditory priming: Implicit and explicit memory for words and voices. *Journal of Experimental Psychology: Learning, Memory and Cognition*, **18**, 915–930.

Tulving, E. (1983). *Elements of episodic memory*. New York: Oxford Univ. Press.


(Received June 20, 1994)

(Revision received December 14, 1994)