Journal of Experimental Psychology:
Learning, Memory, and Cognition
1984, Vol. 10, No. 2, 298-303

Copyright 1984 by the
American Psychological Association, Inc.

Learning Unfamiliar Voices

Gordon E. Legge, Carla Grosmann, and Christina M. Pieper
University of Minnesota (Minneapolis)

Subjects listened to a series of recorded voice samples obtained from unfamiliar speakers and were then given a two-alternative forced-choice recognition test. Recognition performance improved when the voice-sample duration was increased from 6 to 60 s, when the target set size was reduced from 20 to 5 voices, and when slides of faces provided context information. Recognition performance was not significantly different for retention intervals of 15 min and 10 days. For the conditions of our experiments, voice learning was inferior to face learning.

It is commonly accepted that people have the ability to learn unfamiliar voices and the ability to make accurate identifications of familiar voices. However, the factors that affect voice learning and the means by which new voices are encoded in memory are not well understood.

Dating from the seminal work of McGehee (1937, 1944), studies of voice identification and voice memory have usually been concerned with practical aspects of person identification. Several studies of voice learning and memory have been directed at evaluating the reliability of "earwitness" testimony. A good review of these studies is given by Clifford (1980).

Voice-sample duration is likely to affect how well the voice is learned or recognized. Most studies of the effects of voice-sample duration have been concerned with measuring the time required for listeners to recognize familiar voices. It has been shown that samples lasting only a second or so are sufficient for the identification of familiar voices (Bricker & Pruzansky, 1966; Pollack, Pickett, & Sumby, 1954). However, very few studies have examined the effects of voice-sample duration on the learning of unfamiliar voices. Clifford (1980) reported results from two experiments. In the first, groups of subjects heard either one, two, or four sentences spoken by an unfamiliar voice. Subjects then attempted to pick the voice from a six-item voice parade. The three groups showed no significant difference in voice-recognition performance. The second experiment was like the first, except that voice samples of one-half, one, or two sentences were heard, and the subjects were teenagers rather than adults. Recognition scores were slightly higher for the two-sentence group than for the one-half-sentence group.

Because it is likely that a wide enough range of voice-sample durations would have an effect on voice learning, we examined this question further. Groups of subjects were given forced-choice recognition tests following the presentation of voice samples that lasted either 6, 20, 60, or 120 s.

There have been several studies of the effect of retention interval on recognition memory for voices. Carterette and Barneby (1975) found no significant differences in voice-recognition memory for retention intervals of 0, 15, and 45 s. Saslove and Yarmey (1980) required subjects to pick a target voice from a set of five voices, after hearing an 11-s voice sample from the target. They found no significant differences in recognition performance for groups tested immediately after hearing the target voice and groups tested 24 hours later. In two studies, McGehee (1937, 1944) measured recognition memory for voices with retention intervals varying from 1 day to 5 months. Although her two experiments gave slightly different results, there seemed to be little effect of retention interval up to about 1 week, and then a steady decline in performance for longer retention intervals.

Clifford, Rathborn, and Bull (1981) examined the effects of retention interval in experiments in which listeners attempted to pick 2 target voices from a 22-item voice parade.

Requests for reprints should be sent to Gordon E. Legge, Department of Psychology, University of Minnesota, 75 East River Road, Minneapolis, Minnesota 55455.

In their first experiment, retention intervals ranging from 10 min to 130 min had no significant effect on voice recognition. In their second experiment, retention intervals varying from 10 min to 14 days had small, but significant, effects on voice recognition. The results of McGehee (1937, 1944) and Clifford et al. (1981) suggest that voice learning is quite resistant to forgetting over intervals of several days. In these studies, however, listeners were required to remember only 1 or 2 voices. Perhaps rehearsal of such a small target set enabled encoding that was highly resistant to forgetting. Indeed, in one experiment with 5 voices, McGehee (1937) found a decline in voice recognition to chance performance after a retention interval of 1 week. We examined the effects of retention interval, 15 min and 10 days, but with a set of 20 voices to be remembered.

The number of voices to be learned is likely to affect voice learning. Three studies have examined this variable. McGehee (1937) varied the number of target voices from 1 to 5, for groups who were tested after a retention interval of 2 days. She observed no systematic decline in performance until 4 or 5 voices composed the target set. Carterette and Barneby (1975) presented subjects with 2, 3, 4, or 8 target voices. Following a retention interval ranging from 0 to 45 s, subjects were given a recognition test in which they were required to indicate whether a test voice had been one of the original targets. Although there was no effect of retention interval, they found a small decline in performance as the number of target voices increased. In a developmental study, Mann, Diamond, and Carey (1979) presented subjects with 1 or 2 voices to be learned. Subsequent recognition performance was slightly better in the 1-voice case for both children and adults. These studies suggest that the number of voices may be an important determinant of voice learning. We examined this question by measuring voice recognition for groups of subjects who listened to 5, 10, or 20 unfamiliar voices.

Context effects appear to be ubiquitous in studies of memory for verbal material. For example, in word-recognition studies, when context words are paired with target words in a list, subsequent recognition of the targets is improved if the context words are again presented (Thomson, 1972; Tulving & Thomson, 1971). On the other hand, context effects do not seem as important for face recognition. Bower and Karlin (1974) paired context faces with target faces. In subsequent recognition tests, they found no effects of context.

The voice and face of a given individual are stimuli that often accompany one another. We wondered whether faces might provide context information for voice learning. In our final experiment, we paired context faces with target voices. Context effects were evaluated in a subsequent test of voice recognition.

Method

Subjects

A total of 447 subjects formed 26 groups ranging in size from 15 to 23 subjects per group. All subjects were recruited from an introductory psychology course at the University of Minnesota and all were awarded class credit for their participation. For a given experiment, subjects were randomly assigned to conditions. No subject participated in more than one condition.

Stimuli

Our stimuli were recorded voices and slides of faces.

Voices. A library of 46 voices was collected from speakers subject to the following constraints. All speakers were white females, between the ages of 18 and 35, raised in the Midwest, without speech pathology. Apart from these constraints, the speakers were recruited at random from among the acquaintances of the experimenters.

All speakers read the same set of passages from *Grimm's Fairy Tales.* The speakers were instructed to read at their normal conversational level and rate. The spoken passages were recorded on Scotch Grand Master II cassettes under a standard set of recording conditions. These recorded voices were transferred to Maxell 50-60B reel-to-reel tape for presentation during the experiments. The fidelity of the recordings was of good amateur quality. In typical experimental conditions, using these recordings, the experimenters performed with 100% accuracy when naming familiar voices, presented for the shortest durations used in the experiments.

The generality of results from memory-recognition experiments is limited by characteristics of the stimulus set. Undoubtedly, we could have made our voice-recognition task easier if we had deliberately included some speakers with highly unusual voice characteristics, such as accent or speech defect. Such manipulations of the voice set would probably have resulted in differences in the absolute recognition scores measured. However, we were primarily interested in how voice-recognition scores changed with the variation of a number of factors. It seems unlikely to us that qualitative aspects of such changes would depend critically on characteristics of the voice set.

Faces. A library of 50 photographed faces was collected from a Midwest college yearbook, subject to the following

constraints. All the faces belonged to white females having no remarkable facial defects. Black-and-white slides were prepared from Kodak 2565 high-contrast copy film.

Procedure

All experiments were conducted in the same conference room under nearly identical conditions. Groups ranging from 15 to 23 subjects sat around the sides of a U-shaped conference table. The tape recorder was placed at the front of the room with its volume set at a level suitable for comfortable listening. A heavy door excluded virtually all sounds from outside the room.

The following characteristics were common to all the voice-recognition conditions.

A condition consisted of the following two parts separated by a retention interval: an inspection period in which target voices were presented, and a recognition test. Prior to the inspection period, all subjects were told that the purpose of the experiment was to study memory for voices. They were told that they would first hear a series of unfamiliar voices reading the same passage and that their memory for these voices would be tested later with a different passage.

In the inspection period, a series of recorded voices was presented; each voice read an identical passage from *Grimm's Fairy Tales,* separated by 20 s of "dead time" on the tape. After the presentation of a voice, each subject checked a box on the score sheet indicating whether the voice was pleasant or unpleasant. The pleasantness rating was included to aid the subject in concentrating attention on the voices. Orienting tasks of this sort have been shown to result in improved recognition memory for faces (Bower & Karlin, 1974; Patterson & Baddeley, 1977; Warrington & Ackroyd, 1975; Winograd, 1976), but Clifford (1980) reported an unpublished study by Clifford and McCardle in which no effect of orienting tasks was found for voice recognition. Each subject was provided with a written transcript of the spoken passage prior to stimulus presentation. This was done so that the subject would concentrate on the voice rather than the content of the message, even for the first voice in the series.

In the recognition test, a series of forced-choice recognition trials was presented. In a trial, a voice from the series in the inspection period was presented along with a new voice. The two voices were separated by 5 s. All the voices in the recognition test read the same passage, having a nominal duration of 20 s (actual mean = 20.1 s, SD = 2.24 s). The passage in the recognition test was taken from a different story in *Grimm's Fairy Tales.* By changing the content of the passage from inspection period to recognition test, we sought to ensure that recognition judgments were based on characteristics of the voice per se, rather than on idiosyncratic pronunciation of particular words or phrases. An experimenter held a flash card indicating the trial number and whether the first or second voice of the pair was being presented. Subjects were required to indicate on score sheets whether the first or second voice was familiar to them from the inspection period.

The voices in the inspection period were selected at random from the library of 46 voices. An equal number of voices was selected at random from the depleted pool to act as foils. The foils were randomly paired with the targets in the two-alternative forced-choice paradigm. The number of trials in the recognition test was equal to the number of targets in the inspection period, and each of the target voices was presented once in the recognition test. The order of targets in the inspection period, the order of trials in the recognition test, and the order of target and foil within the forced-choice trials were all randomized.

In the first experiment, eight groups participated in a 4 × 2 factorial design to examine the effects of voice-sample duration (6, 20, 60, and 120 s) and retention interval (15 min measured from the termination of the inspection period, and 10 days). (The actual mean durations and percent standard deviations of the 20 spoken passages associated with the four nominal durations are as follows: "6 s" = 5.12 s ± 15%, "20 s" = 24.9 s ± 12.7%, "60 s" = 65.7 s ± 12.5%, and "120 s" = 133.5 s ± 12.5%. In the figures, symbols are plotted at the abscissa values corresponding to these means. For convenience, in the text, we refer to voice-sample durations by their nominal values.) All eight groups heard the same 20 voices in the inspection period in the same order, that order having been randomly determined. The passages associated with the four durations were taken from different stories in *Grimm's Fairy Tales.* All eight groups received an identical recognition test consisting of 20 forced-choice trials.

A face-recognition experiment was conducted with conditions as similar to the voice-recognition experiments as possible. In the inspection period, 20 slides were shown for 6 s each, separated by 20 s. In the recognition test, subjects were given 20 forced-choice recognition trials in which they were required to choose the familiar face from a pair of slides, each presented for 20 s, separated by 5 s. The procedures for selecting and ordering face stimuli were identical to those for the voices. The face-recognition experiment was not strictly comparable to the voice-recognition experiments because identical slides were used in the inspection period and the recognition test. The experiments would have been more comparable if we had changed the pose or expression of the familiar face in the recognition test. However, Patterson and Baddeley (1977) have shown that changes in facial expression and pose have very little effect on face-recognition performance.

In the second experiment, a 3 × 4 factorial design was used to examine the effects of the number of target voices (5, 10, and 20 voices), while once again varying voice-sample duration (6, 20, 60, and 120 s). The retention interval was always 15 min. Other details are like those of the first experiment.

In the third experiment, the context effect of faces was examined. The same four voice-sample durations were again used. This time, the four groups saw faces accompanying the voices. In the inspection period, a face was paired with a voice, the two being presented simultaneously. The subjects were instructed to imagine that the face and voice belonged to the same person. In the recognition test, the face that had been paired with a target voice in the inspection period was projected on the screen for the entire forced-choice trial containing that voice. As a result, the face accompanied both the target and the foil, so the recognition of the face alone could not cue the correct choice. For four other groups, faces accompanied voices in the inspection period only. For all eight groups, the same 20 voices in the same sequence with the same paired faces were used. The retention interval was always 15 min. In other respects, the procedures were identical to those of the first experiment.

Results

Experiment 1: Effects of Voice-Sample Duration and Retention Interval

In Figure 1, the symbols represent the percentage correct for groups plotted as a function of voice-sample duration on a log scale. Fifty percent correct would be expected by chance in the forced-choice paradigm. The filled circles represent data for the four groups having a 15-min retention interval. The open circles represent data for the four groups having a 10-day retention interval.

The most striking result in Figure 1 is the poor voice-recognition performance. Of the eight groups, the best score is only 70%, achieved by the 60-s group with a 15-min retention interval. On the other hand, the scores of the two 6-s groups do not differ significantly from chance on a binomial test of proportions ($p = .05$ criterion). For comparison, we measured face recognition with nearly identical procedures. Twenty slides of faces were presented for 6 s. The percentage correct for the group on the subsequent face-recognition test was 98.5%, confirming results obtained on similar face-recognition tests by Hochberg and Galper (1967). Apparently, for the conditions of our experiments, the learning and recognition of voices were vastly inferior to the learning and recognition of faces.

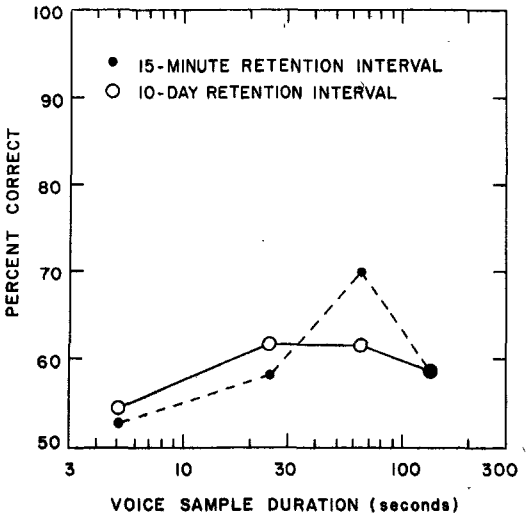

Figure 1. Effects of voice-sample duration and retention interval: Percentage correct for eight groups is plotted as a function of voice-sample duration. (Groups consisted of 15 to 23 subjects, each of whom performed 20 forced-choice trials.)

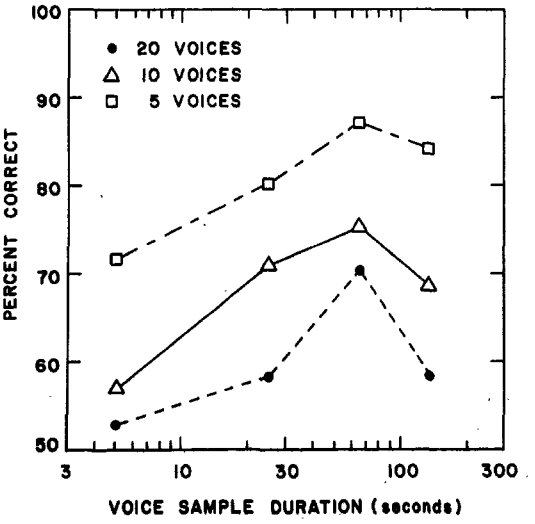

Figure 2. Effect of the number of target voice samples: Percentage correct plotted as a function of voice duration for groups that heard either 5, 10, or 20 target voices.

The data in Figure 1 suggest that increasing voice-sample duration from about 6 s to 60 s leads to improved voice recognition but that recognition is little affected by the change in retention interval from 15 min to 10 days. An analysis of variance (ANOVA) indicated that the effect of voice-sample duration was significant ($p < .01$) but that the effect of retention interval was not significant. The interaction between retention interval and voice-sample duration was significant ($p < .05$). A notable peculiarity in the data of Figure 1 is the decline in performance of the 120-s groups. A t test indicated that the effect is statistically significant ($p < .05$) only for the 15-min retention-interval group. We suspect that this effect was due to boredom. Listening to 20 different voices, each reading the same 120-s passage, is indeed tedious.

Experiment 2: Effect of the Number of Voice Samples

In Figure 2 the filled circles represent data replotted from Figure 1 for the four 15-min retention-interval groups. These groups were presented with 20 voice samples in the inspection period. The squares and triangles represent the overall percentage correct for

groups who heard either 10 voice samples or 5 voice samples, respectively, in the inspection period. The results show that voice learning improves with a reduction in the number of voices to be learned. Separate ANOVAs were conducted, comparing the 20-voice and 10-voice groups, and the 20-voice and 5-voice groups. In both cases, there were significant main effects ($p < .01$) for voice-sample duration and number of voices, but no significant interactions.

Note that the highest score in Figure 2 is only 86.7%, well below 100%. This performance was achieved by a group who heard five target voices, each presented for about 60 s. Despite this rather extensive exposure to a modest number of voices, a considerable number of errors were made in the recognition test after only 15 min.

Experiment 3: Effect of Faces Acting as Context

In Figure 3, the filled circles are group scores replotted from Figure 1. These groups saw no faces. The triangles represent the percentage correct for the four groups for which context information, in the form of faces, was presented in both the inspection period and the recognition test. Clearly, the presence of the faces improved voice-recognition scores substantially.

It is possible that faces aided recognition performance, not by providing cues for retrieval but by providing an orienting task that facilitated encoding of the unfamiliar voices. Appropriate orienting tasks have been shown to facilitate recognition and recall of words (Craik & Tulving, 1975; Hyde & Jenkins, 1973) and recognition of faces (see e.g., Bower & Karlin, 1974). We checked this possibility by repeating our experiment with groups who saw faces during the inspection period only. In Figure 3, the filled squares represent the percentage correct for the four resulting groups. The recognition scores for these groups are much lower than those for the groups who saw faces during both encoding and retrieval of voices. In fact, an ANOVA indicated no significant difference between the groups who saw faces only in the inspection period (filled squares) and the groups who saw no faces at all (filled circles).

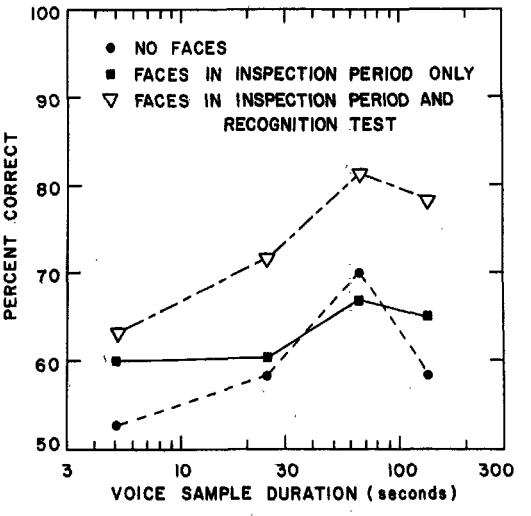

Figure 3. Effect of context: Percentage correct as a function of voice sample duration for 12 groups. (Four groups saw slides of faces accompanying voices both in the inspection period and the recognition test. Four groups saw faces accompanying voices only in the inspection period. Four groups saw no faces.)

These results strongly suggest that faces can provide context information that facilitates memory for voices.

Discussion

The results of the first experiment indicate that the learning of a series of 20 voices is difficult. When the targets were presented for only 6 s each, recognition performance did not differ significantly from chance. However, when the duration was increased from 6 s to 1 min, recognition scores improved. Nevertheless, presentation of as few as 5 voices (Experiment 2), each for 1 min, was followed by errors in forced-choice recognition. On the other hand, the learning that did occur was resistant to forgetting. We found very little difference in recognition performance for retention intervals of 15 min and 10 days. This resistance to forgetting over a period of several days extends similar findings of McGehee (1937, 1944) and Clifford et al. (1981) to the case of 20, rather than 1 or 2, voices.

Our results suggest that voices are more difficult to learn than faces. Perhaps face learning is easier than voice learning because people tend to rely more heavily on faces than voices for identifying people and hence have developed more efficient means for encoding face information. If this is so, it is possible that

people who rely more heavily on voice recognition, namely the blind, would do better in our voice-recognition task. One of the authors has very low vision and relies on voice recognition for the identification of people. It is his experience that voices are indeed difficult to learn. He finds that the task of remembering a set of unfamiliar voices, for example, when meeting with a group of strangers, is a very difficult one. In a comparative study of voice recognition, Bull, Rathborn, and Clifford (1983) found a small, but significant, performance advantage for blind subjects compared with sighted subjects. However, they found no significant relation between recognition accuracy and degree of blindness, age of onset of blindness, or years of blindness.

In Experiment 2, we found that reducing the number of voices to be learned from 20 to 5 led to a substantial improvement in recognition performance. Carterette and Barneby (1975) have suggested that rehearsal may play a role in maintaining a small number of voices in active memory until a recognition test is performed. It is possible that rehearsal of this sort may account for the dependence of our recognition scores on the number of target voices. However, in informal debriefing of subjects following the experiments, none reported rehearsing voices during the retention interval that separated the inspection period from the recognition test.

The results of Experiment 3 surprised us. Faces acting as context can facilitate the recognition of newly learned voices. This result indicates that poor performance in voice recognition is not solely a failure of encoding but is at least partially a failure of retrieval. In some manner face information aids in the retrieval of voice information. It is intriguing to ask whether human faces are unique visual stimuli in their ability to facilitate voice memory. Would pictures of flowers, or alphanumeric characters, or abstract geometrical designs, or chimpanzee faces do as well?

References

Bower, G. H., & Karlin, M. B. (1974). Depth of processing pictures of faces and recognition memory. *Journal of Experimental Psychology, 103,* 751–757.

Bricker, P. D., & Pruzansky, S. (1966). Effects of stimulus content and duration on talker identification. *Journal of the Acoustical Society of America, 40,* 1441–1449.

Bull, R., Rathborn, H., & Clifford, B. R. (1983). The voice-recognition accuracy of blind listeners. *Perception, 12,* 223–226.

Carterette, E. C., & Barneby, A. (1975). Recognition memory for voices. In E. Cohen & G. Nottebohn (Eds.), *Structure and processes in speech perception* (pp. 246–265). New York: Springer.

Clifford, B. R. (1980). Voice identification by human listeners: On earwitness reliability. *Law and Human Behavior, 4,* 373–394.

Clifford, B. R., Rathborn, H., & Bull, R. (1981). The effects of delay on voice recognition accuracy. *Law and Human Behavior, 5,* 201–208.

Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General, 1,* 268–294.

Hochberg, J., & Galper, R. E. (1967). Recognition of faces: I. An exploratory study. *Psychonomic Science, 9,* 619–620.

Hyde, T. S., & Jenkins, J. J. (1973). Recall for words as a function of semantic, graphic and syntactic orienting tasks. *Journal of Verbal Learning and Verbal Behavior, 12,* 471–480.

Mann, V. A., Diamond, R., & Carey, S. (1979). Development of voice recognition: Parallels with face recognition. *Journal of Experimental Child Psychology, 27,* 153–165.

McGehee, F. (1937). The reliability of the identification of the human voice. *Journal of General Psychology, 17,* 249–271.

McGehee, F. (1944). An experimental study of voice recognition. *Journal of General Psychology, 31,* 53–65.

Patterson, K. E., & Baddeley, A. D. (1977). When face recognition fails. *Journal of Experimental Psychology: Human Learning and Memory, 3,* 406–417.

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.

Saslove, H., & Yarmey, A. D. (1980). Long-term auditory memory: Speaker identification. *Journal of Applied Psychology, 65,* 111–116.

Thomson, D. M. (1972). Context effects on recognition memory. *Journal of Verbal Learning and Verbal Behavior, 11,* 497–511.

Tulving, E., & Thomson, D. M. (1971). Retrieval processes in recognition memory: Effects of associative context. *Journal of Experimental Psychology, 87,* 116–124.

Warrington, E. K., & Ackroyd, C. (1975). The effect of orienting task on recognition memory. *Memory & Cognition, 3,* 140–142.

Winograd, E. (1976). Recognition memory for faces following nine different judgments. *Bulletin of the Psychonomic Society, 8,* 419–421.

Received November 20, 1981
Revision received April 27, 1983 ■