Title
Long-term memory for unfamiliar voices.

Journal
The Journal of the Acoustical Society of America, 85(2)

ISSN
0001-4966

Authors
Papcun, G
Kreiman, J
Davis, A

Publication Date
1989-02-01

Peer reviewed

Long-term memory for unfamiliar voices

George Papcun
Los Alamos National Laboratory, P.O. Box 1663, MS B296, Los Alamos, New Mexico 87545

Jody Kreiman
UCLA Phonetics Laboratory, 405 Hilgard, Los Angeles, California 90024 and V.A. Medical Center, Audiology and Speech Pathology (126), Wilshire and Sawtelle Boulevards, Los Angeles, California 90073

Anthony Davis
Department of Linguistics, Stanford University, Stanford, California 94305 and Mitre Corporation, Bedford, Massachusetts 01730

(Received 22 July 1987; accepted for publication 18 October 1988)

From a sample of young male Californians, ten speakers were selected whose voices were approximately normally distributed with respect to the "easy-to-remember" versus "hard-to-remember" judgments of a group of raters. A separate group of listeners each heard one of the voices, and, after delays of 1, 2, or 4 weeks, tried to identify the voice they had heard, using an open-set, independent-judgment task. Distributions of the results did not differ from the distributions expected under the hypothesis of independent judgments. For both "heard previously" and "not heard previously" responses, there was a trend toward increasing accuracy as a function of increasing listener certainty. Overall, heard previously responses were less accurate than not heard previously responses. For heard previously responses, there was a trend toward decreasing accuracy as a function of delay between hearing a voice and trying to identify it. Information-theoretic analysis showed loss of information as a function of delay and provided means to quantify the effects of patterns of voice confusability. Signal-detection analysis revealed the similarity of results from diverse experimental paradigms. A "prototype" model is advanced to explain the fact that certain voices are preferentially selected as having been heard previously. The model also unites several previously unconnected findings in the literature on voice recognition and makes testable predictions.

PACS numbers: 43.71.Bp

INTRODUCTION

This study addresses the question of how well people remember unfamiliar voices after delays of 1, 2, and 4 weeks and examines the processes underlying memory for voices. These issues are of both practical and theoretical importance since their answers bear on matters including the specifications required for speech storage and transmission systems, the validity of legal testimony involving identification by voice, and the general problem of auditory pattern recognition.

Most previous studies of long-term memory for unfamiliar voices have used closed-set, multiple-choice formats (e.g., McGehee, 1937, 1944; Clifford *et al.*, 1981; Saslove and Yarmey, 1980; Legge *et al.*, 1984). In these tasks, listeners hear a target voice that they later attempt to select from among a set of voices in which they are truthfully informed that it appears once and only once. For example, in the earliest of these studies (McGehee, 1937, 1944), listeners attempted to select a single target voice from a set of five male voices after delays that ranged from 1 day–5 months. Recognition scores declined from 83% after 1 day to 80.8% after 1 week, 68.5% after 2 weeks, 57% after 1 month, and to 13% after 5 months. McGehee provided no tests of the statistical significance of the decline, and subsequent studies have generally failed to show statistically significant differences between delay conditions in listeners' ability to identify voices over the delays studied here (see Clifford, 1980, for a review). In the 1937 study, McGehee used listener groups of differing sizes at the various delays, but she does not specify which groups were used at which delays. Moreover, she states that in two cases there was a "break in discipline" with "a spontaneous acclamation by the majority of students when they heard the repeating voice" (p. 259), but she does not specify in which cases or at which delays this occurred.

Thompson (1985) used male voices in a six-voice lineup task in which listeners rated each voice as to whether it was the voice they had heard 1 week previously. They could also respond that the voice heard previously was not in the lineup or that they were not sure whether it was in the lineup. However, the listeners were not given the option of saying the voice heard previously was in the lineup more than once. Thus, from the viewpoint of the listeners, the experiment was an open-set task, but not an independent-judgment task. Such a task can be considered an open-set, multiple-choice task with a decision threshold imposed by the listener. The results were 62.1% correct identifications, 22.1% incorrect identifications, and 15.8% "not in lineup" or "not sure if in lineup" responses.

The present study used an open-set, independent-judgment recognition task in which listeners each tried to remember a single voice. In the recognition phase of the experiment, the listeners were told that the voice that they heard previously might appear once, more than once, or not at all. They were, therefore, to make each judgment independently of all others. This task is more veridical to most realistic

 0001-4966/89/020913-13$00.80

situations than closed-set formats; moreover, as we shall demonstrate below, it yielded data that ultimately proved revealing of the processes underlying listeners' judgments.

Clifford (1980) states that "experimentation in this area is characterized by the lack of generalizability, by the lack of comprehensiveness; by the lack of a sound theoretical or explanatory perspective" (for similar sentiments, see also Bricker and Pruzansky, 1976). It is our purpose to remedy these deficiencies by providing appropriate data, analysis, and theory, and to do so in the context of an experiment that allows us to interpret our results in terms of realistic situations.

I. METHOD

A. Initial speaker selection

Twenty-two male speakers were recruited by means of an advertisement in the UCLA campus newspaper. Speakers ranged from 19–31 years of age. All had lived in California at least since adolescence and were without regional accent other than typical of California, as judged by the authors.

B. Voice samplers and recording procedures

The speakers were recorded while making telephone survey calls. This technique allowed the recording of a controlled text in the context of interpersonal interaction and resulted in natural-sounding speech samples. The topic of the survey was attitudes toward crime. It included statements and questions of varying lengths and structures.

Recordings were made in a quiet office on a Uher 4200 reel-to-reel tape recorder[1] on low print-through tape at 7½ in. per second. The interviewers were not recorded through the telephone; instead, they were recorded with a high-quality dynamic microphone that was attached to the telephone mouthpiece. Good-quality recordings were thus obtained while allowing the interviewers to engage in normal telephone conversations. Only the interviewer's voice, and not the voice of the interviewee, was recorded.

In order to sample within, as well as between, speaker variability in voice quality and speech mannerisms (see Stevens, 1972; Nolan, 1983), each interviewer made four survey calls, two in each of two sessions at least 1 week apart. Two of the four survey calls—one from each recording session—were selected for each speaker. The calls were selected on the basis of naturalness, fluency, conformity to the text, and lack of extraneous comments. All interviewers occasionally strayed from the text of the survey to comment on the answers, to ask further questions, and so on. To maintain constancy of the material presented to the listeners, these digressions, as well as lengthy pauses and excessive numbers of filled pauses, were edited out of the two calls selected. However, hesitations, false starts, disfluencies, most filled pauses, minor unfilled pauses, and minor rewordings were not edited. The edited recordings were transferred to cassettes for convenience in playing them to listeners. The edited recordings lasted an average of 1.58 min each (s.d. = 0.13 min).

C. Final voice selection: "Easy-to-remember" versus "hard-to-remember" ratings

Pilot tests suggested that listeners can attend to only about ten voices in a single listening session. Therefore, seven speakers who strayed too far from the script or whose speech mannerisms did not match those of the other speakers, in the judgment of the experimenters, were eliminated from further consideration. The following procedure was applied to select among the remaining 15 voices.

Five groups of ten listeners were asked to rate, for each of the 15 voices, how easy or hard they thought the voice would be to remember. The raters were told that all speakers were young male Californians. They responded on a seven-point scale, with 1 meaning "very easy to remember" and 7 meaning "very hard to remember." Each group of ten raters heard the voices in a different random order. Only the first half of the survey call was played, since pilot tests suggested that this was sufficient for the judgment required. Raters also heard one practice voice not used later in the test to familiarize them with the text and procedures.

Two tests of the reliability of agreement among raters were performed. First, raters were divided at random into two groups, and the mean rating for each voice was calculated for each group. These mean ratings were significantly correlated for the two groups; Pearson's $r = 0.73$, with 13 df, $p < 0.001$. Additionally, to examine the effect of presentation order on ratings, a two-way (voices × presentation order) fixed effects ANOVA with repeated measures on voices was performed. The effect of voices on ratings was significant [$F(14, 630) = 9.408, p < 0.01$], as was the voice × presentation order interaction; $F(56, 630) = 2.433$, $p < 0.01$. No other effect, including the main effect of presentation order, was significant at the 0.05 level.

Ratings for each voice were totaled across presentation orders and standardized. A plot was made of standardized scores versus the rank order of the voices on the easy-to-remember versus hard-to-remember ratings, and obvious outliers were eliminated. Two voices that were described as accented by an appreciable number of raters were also eliminated. Finally, those ten voices that most nearly approximated a normal distribution on hard-to-remember versus easy-to-remember ratings were selected by reference to a cumulative seminormal plot. Thus the voices that were selected are approximately normally distributed with respect to how difficult listeners believe they would be to remember, but are otherwise unspecified with respect to voice characteristics except that they are closely controlled for regional dialect, age, and sex, as described above.

On the basis of the preceding results, three voices were selected as target voices: speaker 2, whose voice was judged by the combined rankings of all judges as next to easiest to remember; speaker 9, whose voice was judged next to hardest to remember; and speaker 5, whose voice was judged intermediate in difficulty.

D. The experimental design

A total of 90 listeners, all native speakers of English, were divided randomly into three groups of 30. Each of the three target voices was played to one of the three groups of

listeners; each group heard only one target voice. Groups of five or fewer listeners heard the recordings on a good-quality cassette player (Marantz model PMD 360) in a quiet room. The listeners were told that they would hear the voice of a young male Californian, and they were asked to pay very close attention to the voice, since they would later hear a group of voices and would have to decide if the presented voice was in it or not, and if it was, to identify it. Listeners then heard one complete survey call taken from the speaker's first recording session.

For each target voice group, ten listeners returned after 1 week, ten listeners returned after 2 weeks, and ten listeners returned after 4 weeks. When they returned, the listeners were informed that they would hear ten recordings of young male Californians, all taking the same telephone survey that they had heard previously. They were told that the voice they heard at the previous session (the target voice) might appear once, more than once, or not at all. They were told that, if the target appeared, they would hear a different recording of it than they had previously heard.

In reality, listeners heard each of the ten voices described above, including the target voice, once only. The voices were played in one of two orders: the second was the reverse of the first, with the exception that the target voice appeared in seventh position in both orders. Full survey calls were played for every trial. The calls made at the speakers' second recording session were used in this phase of the experiment, so the listeners did not hear the same recording twice.

At each of the three delays, three groups of ten listeners each heard the voice that they had heard previously. They also heard nine voices they had not heard previously. Thus the experiment yielded 300 data points at each delay and 900 data points altogether.

For each voice, listeners were to indicate whether or not it was the voice they had heard at the first listening session. They were also to indicate how certain they were that their answer was correct by using a scale of 1 to 5, where 1 meant "extremely certain the answer is correct," and 5 meant "extremely uncertain the answer is correct." If they thought the probe voice was different from the previously heard voice, they were asked to use a five-point scale to indicate how similar the probe voice was to the previously heard voice. The similarity judgment data will not be reported in this article. A separate answer sheet was provided for each voice. It was turned over before the next voice, and listeners were not allowed to turn back.

II. RESULTS

A. Correct identifications and false identifications

The number of correct identifications and false identifications for each of the target voices when it was a target at each delay interval is shown in Table I.

As mentioned above, listeners were asked to rate their confidence in each of their answers on a scale of 1 to 5. We will mark those certainty judgments for which listeners claimed they had not heard the voice previously with minus signs. Thus combining listeners' confidence ratings with their heard previously/not heard previously responses yields a ten-point scale ranging from certainty + 1 heard previously (i.e., extremely certain that the voice is the voice heard previously) to certainty − 1 not heard previously (i.e., extremely certain that the voice is *not* the voice heard previously).

B. Independence of successive judgments

Listeners were instructed that the voice they heard at the first listening session might occur once, many times, or not at all during the second listening session; hence, by implication, they were to make each judgment independently of all others. Nonetheless, the experimental design leaves open the possibility that listeners' judgments might be influenced by their previous judgments. For example, one might expect that listeners who had already identified one voice as the target (whether correctly or not) would be less likely to claim that any of the remaining voices was also the target. As well as being of interest in its own right, this issue is of importance because the information-theoretic and signal-detection analyses we apply below depend on the assumption of independence.

Therefore, to test the hypothesis of independence, we calculated the distributions of the numbers of listeners who would be expected to make various numbers of false identifications if each judgment was made independently of the others, and then compared the theoretically expected distributions with the observed distributions. The derivation of the

TABLE I. Numbers of correct and false identifications engendered by each target voice when it was a target at each delay.

		Target voice			
Correct identifications		Easy	Medium	Hard	All voices
Delay (weeks)	One	6	7	7	20
	Two	8	7	6	21
	Four	5	6	6	17
	All delays	19	20	19	
False identifications					
Delay (weeks)	One	7	5	4	16
	Two	9	11	7	27
	Four	12	7	6	25
	All delays	28	23	17	

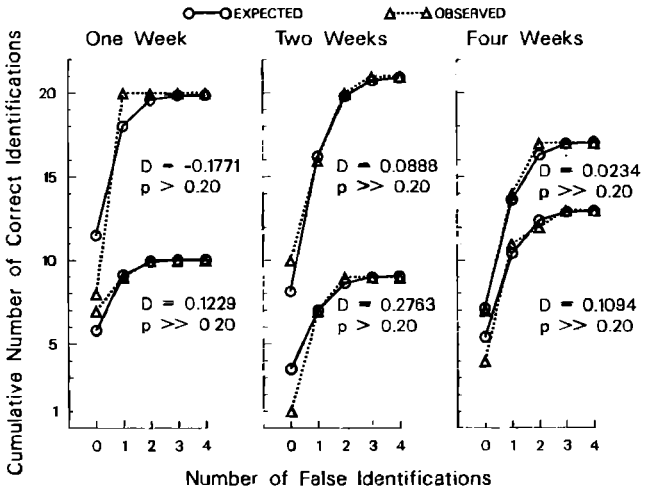

FIG. 1. Cumulative distributions of numbers of listeners with 0, 1, 2, and 3 errors. The theoretically expected curves are shown with solid lines, and the observed results are shown with broken lines.

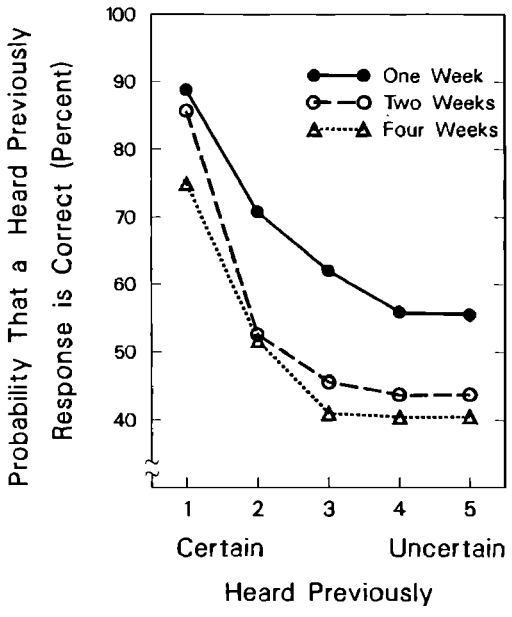

FIG. 2. Probability that a heard previously response at each certainty level is correct.

theoretically expected distributions is given in the Appendix.

The theoretically expected distributions are shown alongside the observed distributions in Fig. 1, which also shows the Kolmogorov–Smirnoff statistic for ordered categories, D, and its associated probabilities (Siegel, 1956, pp. 127–136). Overall, the empirical data are extremely similar to the results predicted under the assumption of independent judgments. In no case is the difference significant at the 0.20 level; therefore, there is no evidence that the listeners did not make each of their judgments independently, as they had been instructed to do.

C. Probabilities that responses are correct or incorrect

An issue of immediate interest is the probability that a listener's response at a specified level of certainty is correct or incorrect. This is found by applying Bayes' rule:

$$P(x_i|y_j) = [P(y_j|x_i)][P(x_i)]/P(y_j),$$

where x_1 is the voice that was heard previously, x_2 are the other voices, y_1 is the response heard previously, and y_2 is the response not heard previously. In applying Bayes' rule to these data, we treat the certainty levels as sequences of partitions. Hence, at certainty level + 1, we include only the data for certainty + 1 (extremely certain heard previously); at certainty level + 2, we include the data for levels + 1 and + 2, and so on. For this analysis, we sum the heard previously responses left to right, from + 1, "extremely certain heard previously," through + 5, "uncertain heard previously," and we sum the "not heard previously" responses right to left, from certainty − 1, "extremely certain not heard previously," through − 5 "uncertain not heard previously." By summing the data this way, we assess the heard previously and the not heard previously judgments separately, each as a function of certainty. In effect, the certainty levels are interpreted as sequences of thresholds, with each subsequent threshold less stringent than its predecessor. According to this interpretation, any judgment that passes a relatively stringent threshold (i.e., a more certain judgment) would have passed any laxer threshold as well, which, in turn, justifies subsuming the data from more stringent thresholds with the data from laxer ones. This line of reasoning is typically used in signal-detection analyses (see Sec. I E) and has been empirically justified by numerous experiments (see Swets, 1964, for review).

The results of applying Bayes' rule to the data in this manner are shown in Figs. 2 and 3. We evaluate the apparent trend toward increasing accuracy as a function of increasing

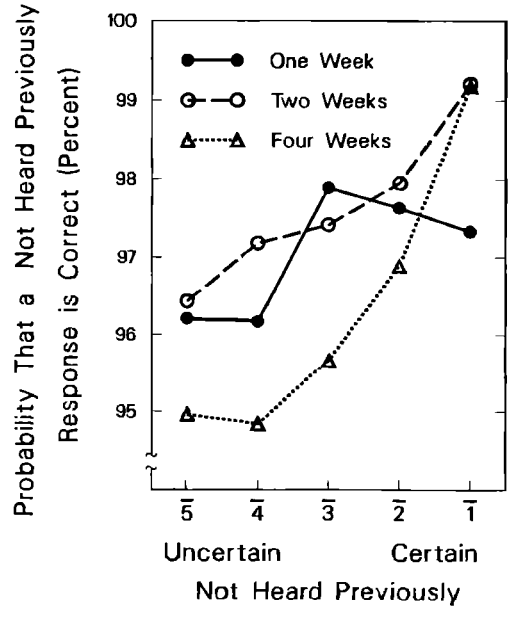

FIG. 3. Probability that a not heard previously response at each certainty level is correct.

certainty with the nonparametric approximate test for trend of related samples, an extension to Page's L (Meddis, 1984, pp. 224–229). For the heard previously responses, $L = 164$; z corrected for ties = 3.40; $p < 0.001$. For the not heard previously responses, $L = 159$; $z = 2.77$; $p < 0.01$.

At all delays and all certainty levels, heard previously responses were less accurate than not heard previously responses. (Note the difference in the ordinate scales of Figs. 2 and 3.) The significance of the difference in accuracy overall is tested by

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}\hat{q}(n_1)^{-1} + (n_2)^{-1}}}$$

(Snedecor and Cochran, 1967, pp. 220–221). We find that $z = 16.35$; $p \ll 0.001$.

Evaluating the statistical significance of the apparent trend as a function of delay presents certain problems. One problem is that measures based solely on the responses that fall into the various categories do not take into account the number of instances that gave rise to the responses. Another problem is that the data are not independent because a fixed number of decisions are allocated to a fixed number of categories and also because the data are accumulated. No analytical statistical methods known to us are appropriate to deal with these issues.

Therefore, we analyzed the data with a bootstrap simulation. Briefly, the bootstrap method involves developing a computer model of the experiment and sampling data randomly with replacement. Following a large number of runs, the result of the simulation can be observed for parameters of interest (Efron, 1979a,b; Diaconis and Efron, 1983; for a discussion of qualms about the bootstrap method as applied to determining confidence intervals, see Schenker, 1985).

To examine the significance of the apparent trend over time, the bootstrap simulation was run with flat input data, i.e., with data such that three identifications and 30 rejections were ascribed to each partition at each delay. For the heard previously results, we asked how frequently flat input data would yield the result that the three curves are ranked in descending order as a function of delay at all five points, as was the case for the experimental results. If the points were independent, the probability of the observed ordering would be one-sixth to the fifth power, or 0.00013; however, because the data are summed, the probability that the points at each succeeding partition will be in the observed order is influenced by the order of the points at the previous partition. The bootstrap model incorporated the effect of summation. With one million bootstrap iterations, the probability that all five points would lie in the observed order was found to be 0.014 (see footnote 2). For the not heard previously responses, only the curves for the two longer delays are in the predicted order. These curves are defined by relatively few responses, especially at the 1-week delay, and are therefore relatively unstable. Consequently, we make no claim with respect to the trend over time for the not heard previously results.

D. Information-theoretic interpretation

In an information-theoretic interpretation of this experiment, the listeners' memories are considered to be a communications channel that can lose the capacity to transmit information over time. More precisely, the communications channel is considered to be the listeners' ability to discriminate their memory of the voice heard previously from their perceptions of other voices, within the context of the experimental task. In this analysis, the input to the communications channel consists of the voices presented at the second listening session. The listeners' responses constitute the output of the channel. The target voice defines the correct output, and thereby presumably affects the transmission characteristics of the channel. The voices presented at the second listening session constitute a set of symbols that are transmitted to the output. To the extent that their responses do not consistently reflect the input, the listeners are said to be losing information. An information-theoretic analysis differs from the other analyses we present in that it incorporates the effects of the experimental task as part of the channel and can reveal the effects of differing degrees of confusability among the nontarget voices. We show in this section that these effects are more pronounced when listeners use laxer criteria for selecting the voice heard previously, and when results with different nontarget voices are considered separately. For expositions of information theory and derivations of the formulas used below, see McEliece (1977) and Attneave (1959).

In the present experiment, there are two ways to view the set of input voices. One is to consider the voice that was heard previously (the target voice) to be a symbol having the value HEARD PREVIOUSLY, while all other voices are symbols having the value NOT HEARD PREVIOUSLY, that is to say that there are just two input symbols with unequal distributions. Alternatively, we may treat each of the ten voices presented at the second listening session as a different stimulus; according to this view, there are ten different input symbols with equal distributions. As we shall see, the first approach is simpler, but the second approach discloses phenomena not disclosed by the first.

In this experiment, the distribution of the input symbols, and therefore the amount of information in the source, is set by the design of the second listening session, where the target voice is heard once and other voices are heard nine times. First, we consider the analysis with two input symbols. Designating the voice that was heard previously as x_1, and designating any of the voices that were not heard previously as x_2; $P(x_1) = 0.1$, and $P(x_2) = 0.9$. From the distribution of the input symbols, we calculate the information of the source as

$$H(X) = \sum P(x_i) \log_2 \frac{1}{P(x_i)} = 0.469 \text{ bits.}$$

This quantity is the same for all delay conditions. Because this calculation uses base 2 logarithms, the result represents the average number of binary distinctions (bits) needed to specify the distribution of the source.

It is possible to compute the amount of information that observing the output tells about the input. This measure is known as the mutual information. When the mutual information is normalized as a proportion of the information in the source, a measure of the relative information transmission (RIT) is produced. The RIT is found by the following:

$$\text{RIT} = \sum_j \sum_i P(x_i)P(y_j|x_i)\log_2\left(\frac{1}{P(x_i|y_i)}\right)\Big/H(X).$$

Thus the RIT is a measure of the categories actually distinguishable at the output relative to the categories potentially distinguishable at the input. The relative information transmission for each of the three delays is shown in Fig. 4. The standard deviations about each point were computed by means of a bootstrap simulation with one million iterations.

At each bootstrap iteration, the slope of the regression of relative information transmission on delay was calculated. By this calculation, we find the probability that the linear decline is not negative to be less than 0.032, thus indicating that relative information transmission declines as the delay between voice presentation and identification increases.

Figure 5 shows the relative information transmission for each of the target voices. Note that the results do not appear to be in the direction predicted by the group of listeners who rated the voices as easy or hard to remember. We will return to this issue below.

We now consider the experiment as a task of classifying ten input stimuli (rather than two input stimuli, as above) into two response categories. Since at the second listening session each of the ten voices was presented once to each listener, the amount of information in the source is

$$H(X) = \sum_{i=1}^{10} p(x_i)\log_2\frac{1}{p(x_i)} = \log_2 10$$
$$= 3.3219 \text{ bits.}$$

Ideal performance in this experiment occurs when the probability of responding heard previously is 1 to the voice that was, in fact, heard previously and 0 to all other voices, that is, when the nine nontarget voices are collapsed into a single response category. The relative information transmission is then 14.12%.

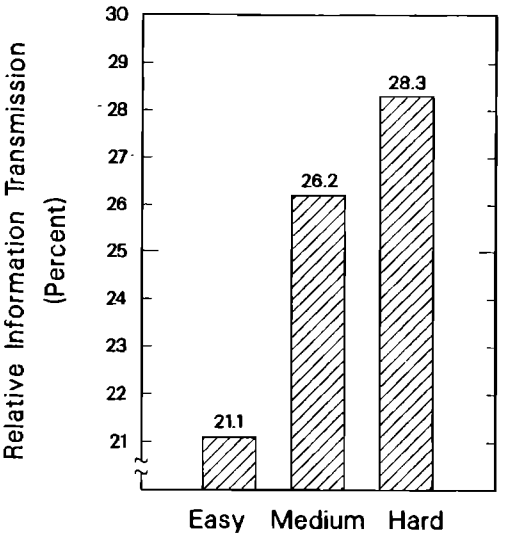

FIG. 5. Relative information transmission for each of the target voices.

Minimum information loss for this system does not, however, correspond to ideal performance in the sense just defined. Rather, minimum information loss occurs when the inputs are apportioned among the outputs in such a way that they are maximally distinguished within the limits set by the task. In this system, this would occur when listeners said heard previously to half of the voices presented and not heard previously to the other half. In an information-theoretic analysis, the distribution of responses affects the amount of information that is transmitted. The task constrains the way in which the answers can be distributed, thereby limiting the amount of information that can be transmitted. Therefore, the task must be regarded as part of the channel.

Figure 6 shows the values of relative information transmission at each delay interval for the target voices combined

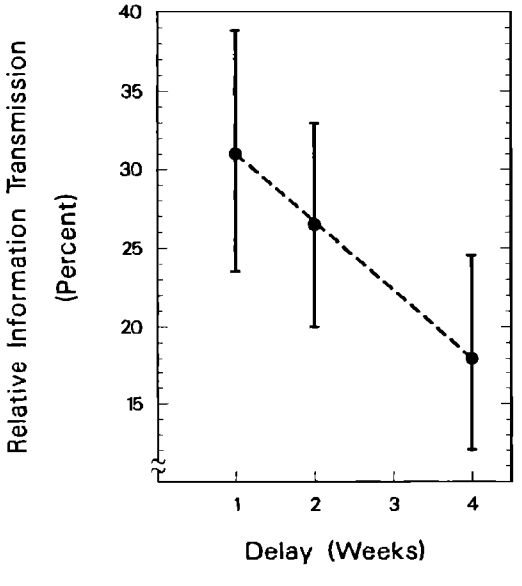

FIG. 4. Relative information transmission at each delay. Standard deviations are shown as vertical bars at each point. The regression of relative information transmission on delay weighted by the inverse of the variance at each point is shown by the diagonal broken line.

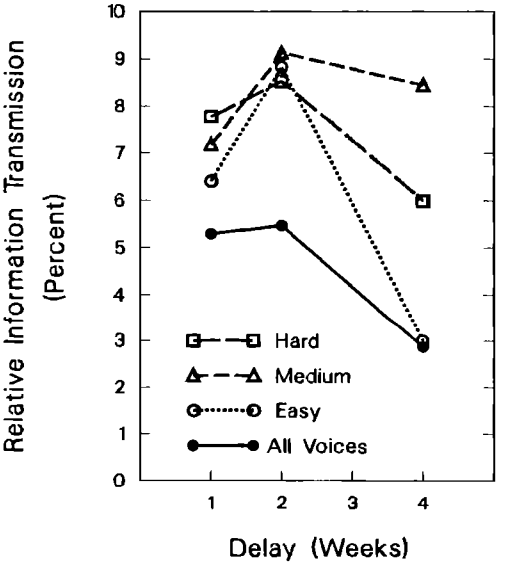

FIG. 6. Relative information transmission as calculated according to the ten-input analysis.

and for each of the target voices separately. Note that the relative information transmission values for each target voice considered separately exceed those for the combined data, in contradiction to the naive assumption that the combined results should be the mean of the separate results. Also note that there is a rise in relative information transmission values at the 2-week delay.

These two results have the same underlying explanation, viz., the nonuniform distribution of false identifications among the nontarget voices. Table II shows the numbers of heard previously responses to each voice when it was not a target, separated both by target and by delay. Notice that there is a nonuniform association between the nontarget voices and the false identifications elicited by each target voice. Thus listeners are effectively classifying a given target voice and several other voices into one category, and classifying the remaining voices into another category. Combining the data from all target voices, however, results in a more random distribution of false identifications, and, consequently, in lower relative information transmission. Listeners in the 2-week delay produced the largest number of false identifications, thus enhancing the effects of the nonuniform distribution of false identifications on information transmission. The fact that the relative information transmission for individual target voices is greater than the relative information transmission for the combined data results from and serves to quantify the fact that some subsets of voices are more confusable than others (Thompson, 1985; Rosenberg, 1973; Bricker and Pruzansky, 1976). These results also highlight the sense in which a false identification may be informative even though it is mistaken. From a different perspective, they emphasize the danger of misidentification, albeit informative.

E. Signal-detection interpretation

In a signal-detection interpretation of this experiment, we consider the voice that was heard at the first listening session to be a signal that the listener is trying to detect. All voices other than the signal voice are considered noise, and detection consists of distinguishing the signal voice from noise voices. As above, we treat the certainty levels as a basis for cumulative partitions of the data, and interpret the partitions as a sequence of thresholds. In a signal-detection analysis, the cumulative partitions of the data are interpreted as a single sequence of thresholds ranging from the most stringent criterion for the acceptance of a token as a signal (+ 1 certain heard previously) through the laxest criterion for acceptance of a token as a signal (− 1 certain not heard previously). Thus, unlike the Bayesian analysis, the signal-detection analysis does not distinguish possibly differing listener strategies for heard previously versus not heard previously responses. On the other hand, given assumptions about the distributions underlying listener's responses, signal-detection analysis makes it possible to draw inferences about responses under varying circumstances, as we shall demonstrate below.

Receiver operating characteristic (ROC) curves characterize the signal-detection capabilities of the receiver—in this case, the listeners. An ROC curve shows the cumulative probability $P(y_1|x_1)$ (the probability that listeners correctly identified a voice that they heard previously) on the ordinate against the cumulative probability $P(y_1|x_2)$ (the probability that listeners falsely identified a voice as having been heard previously) on the abscissa. We coalesced the data into three partitions, the minimum number necessary to perform an ROC analysis, thus achieving the most stable estimates of detection sensitivity at each delay at the expense of information about the effects of the individual certainty levels.[3] We will examine the individual certainty levels below. Using a program by Dorfman and Alf (1969, reprinted in Swets and Pickett, 1982) for maximum likelihood analysis of binormal ROC curve data, we calculated the ROC curves shown in Fig. 7.

One measure of detection sensitivity is the area beneath the binormal ROC curve, with greater area representing greater sensitivity. This measure sums over all thresholds, and in that sense is independent of any of them individually. The areas beneath the binormal ROC curves are shown in Fig. 8. The standard deviations of the areas were also calculated by the Dorfman and Alf program. We assessed the hypothesis of a linear decline by the method of the contrast coefficients adjusted for the unequal intervals of delay (Keppel, 1982, pp. 629–634); $t = 1.79$, with 87 df and $p < 0.05$ (see footnote 4).

Other measures of detection sensitivity are d' and d'_e (Green and Swets, 1974). In general, d'_e is the preferred measure because it takes into account the variances of both the signal and the noise distributions, but, in the case of a forced-choice experiment, that information is not available,

TABLE II. The numbers of heard previously responses to nontarget voices for each target and at each delay.

		Nontarget voices in order from easy to hard									
		1	2	3	4	5	6	7	8	9	10
Target voices	Easy	0	x	2	0	3	1	2	9	5	6
	Medium	1	1	1	1	x	1	3	2	13	0
	Hard	0	0	1	1	8	1	2	3	x	1
Delay (weeks)	One	0	0	1	1	1	0	4	4	4	1
	Two	0	1	1	1	6	1	2	6	7	2
	Four	1	0	2	0	4	2	1	4	7	4

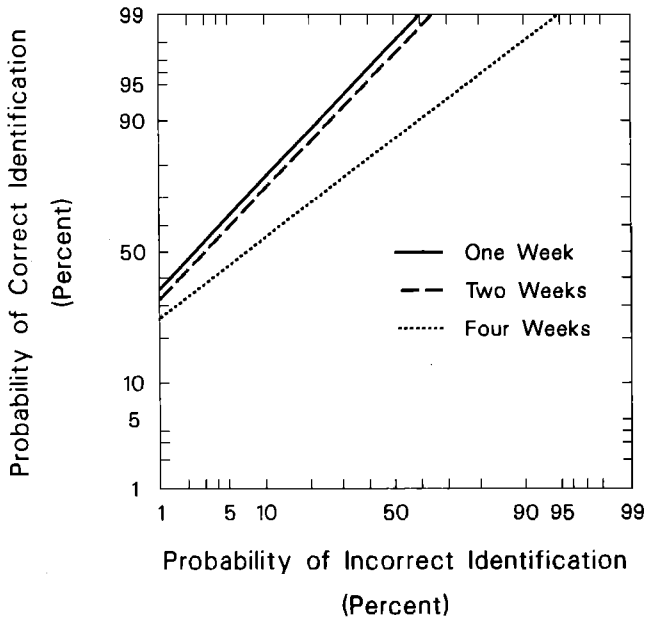

FIG. 7. Binormal receiver operating characteristic curves at each delay.

so d' must be used. For our experiment, at 1 week, $d'_e = 2.05$, at 2 weeks, $d'_e = 1.93$, and at 4 weeks, $d'_e = 1.27$.

As we mentioned in the Introduction, McGehee (1937, 1944) tested recognition of an unfamiliar voice with a closed-set, multiple-choice task in which listeners heard a single voice and later tried to select it from among a group of five voices. We emphasize that the open-set, independent-judgment task we used differs considerably from the task used by McGehee. It has been shown by Birdsall and Peterson (1954) and by Green and Birdsall (1964) that it is possible to derive d' for m orthogonal response alternatives on the basis of the largest of m drawings from a normal deviate (Tippett's Distribution; see Tippett, 1925). Birdsall and Peterson provide nomograms for d' as a function of the number of alternatives and the percentage of correct responses. With these nomograms, we determined d' for McGehee's 1937 data. [McGehee's (1944) study produced results congruent with those of the 1937 study but did not include the same delays as our study.] At 1 week, $d' = 2.04$, at 2 weeks, $d' = 1.54$, and, at 4 weeks, $d' = 1.20$. Standard deviations for McGehee's results were calculated on the basis of binomial proportions, which were then interpreted in terms of d' from Green and Birdsall's nomograms. McGehee used different sized groups of listeners for different delays and, unfortunately, does not specify which group was used for which delay. The smallest group was 34 listeners, so we base our estimates of standard deviations for her data on that number. Therefore, our estimates of standard deviations for her data are liberal in the sense that larger groups would yield smaller estimates of the standard deviations.

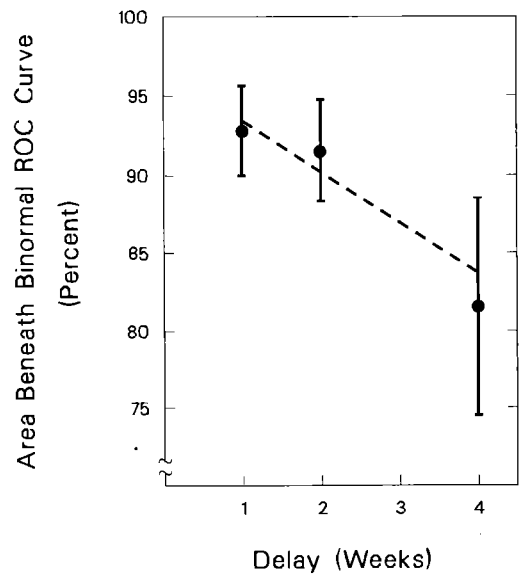

FIG. 8. The area beneath the binormal receiver operating characteristic curve at each delay. Standard deviations are shown as vertical bars at each point. The regression of the area beneath the receiver operating characteristic curve on delay weighted by the inverse of the variance at each point is shown by the diagonal broken line.

Also mentioned in the Introduction was Thompson (1985), who used a task in which listeners were to choose which voice of a six-voice lineup was the voice heard previously but were also allowed to respond that it was not in the lineup or that they were not sure whether it was in the lineup. In signal-detection terms, this task can be considered an open-set, multiple-choice task in which listeners set a threshold according to which they decide whether or not to choose any of the voices. Papcun *et al.* (1983) show that it is possible to analyze the data from such a task and predict the results that would have been obtained without a threshold,[5] i.e., had listeners not been allowed to say "not in lineup" or "not sure if in lineup." The derived results can be analyzed as though they were the results of a closed-set, multiple-choice task. Applying this analysis to Thompson's data for male voices, we estimate that, without a threshold, listeners would have achieved 74% correct identifications. Green and Birdsall's nomograms show that for a six-alternative multiple-choice task, at 74% correct identifications, $d' = 1.97$. Standard deviations were calculated on the basis of binomial proportions, and then interpreted in terms of d' from the nomograms.

In Fig. 9, we show d'_e for our results and d' for McGehee's and Thompson's results. Considering that the experiments whose results are shown in Fig. 9 were separated by almost 50 years, and that three different experimental paradigms were used, the correspondence among the results is striking, a fact that we take as confirmation not only of the results, but also of the analyses we have applied. A further implication of this finding is that it is possible to make valid predictions about rates of correct and false identifications and eliminations across such diverse paradigms.

We now examine the ROC curve points representing the individual certainty levels. A consistent and interpretable pattern emerges when we consider the points defined by the first five certainty levels, a region of particular interest because these points correspond to listener's claims that the voice in question was heard previously. The first five points

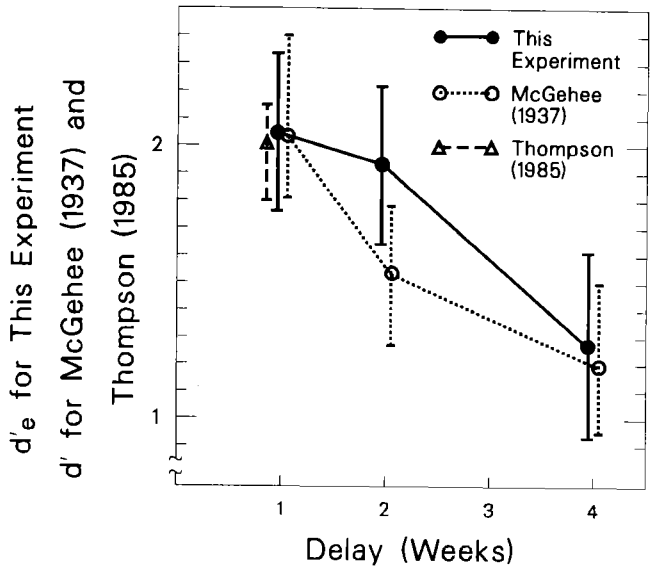

FIG. 9. Values of d'_e for this experiment and d' for McGehee (1937) and Thompson (1985).

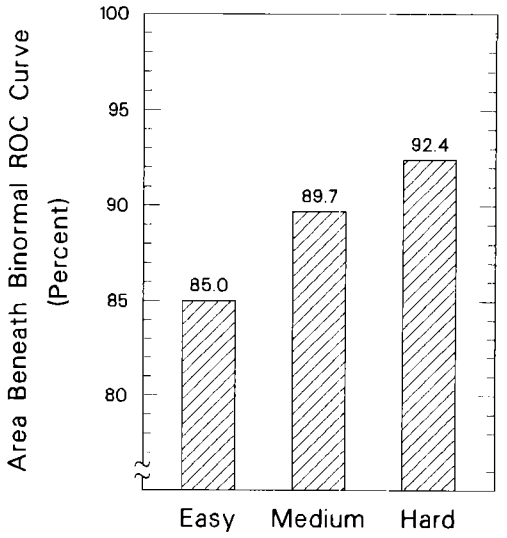

FIG. 11. The area beneath the binormal receiver operating characteristic curve for each of the target voices.

of the ROC curve are shown with a magnified horizontal scale in Fig. 10. The curves lie in the expected relationship to one another, that is, generally beneath each other as the delays increase, and points for corresponding certainty levels are higher and farther to the right for the 2-week curve than for the 1-week curve. This implies that, within the region of the first five cetainty levels, the 2-week results are biased toward more heard previously responses than the 1-week results.

The areas beneath the binormal ROC curves for each of the three target voices, with the data collapsed over the three delay intervals, are shown in Fig. 11.

The pattern of areas beneath the curves may initially seem surprising in view of the predictions made by the independent group of listeners who rated the voices as to whether they would be easy or hard to remember. However, in the following section, we will show that the pattern of the areas beneath the ROC curves is, in fact, consistent with those ratings.

F. Easy-to-remember versus hard-to-remember ratings: A "prototype" interpretation

We turn now to a more detailed analysis of the voices in terms of their easy-to-remember versus hard-to-remember ratings, and we present an interpretation of the analysis.

Figure 12 shows the numbers of correct identifications and false identifications that were engendered by each of the

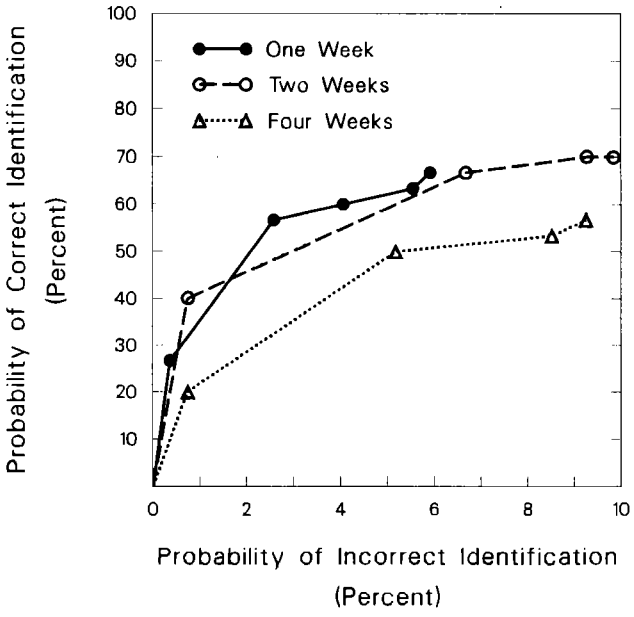

FIG. 10. Receiver operating characteristic curves for those points at which listeners claimed the probe voice was heard previously.

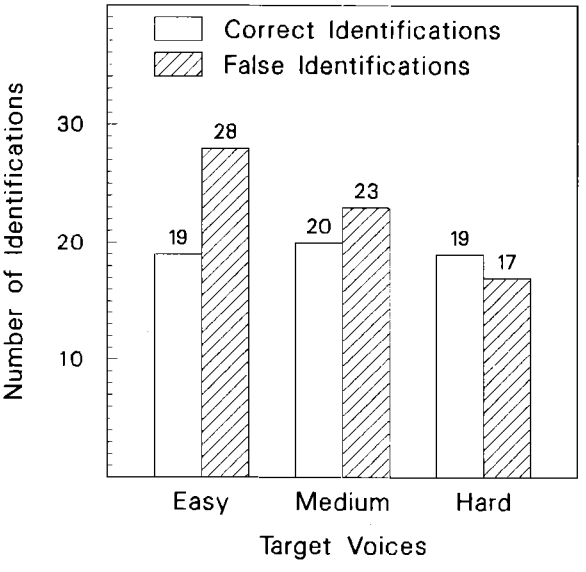

FIG. 12. Numbers of correct and false identifications engendered by each target voice.

target voices when it was a target, i.e., when it was the voice that had been presented at the first listening session. Thus the false identifications shown in Fig. 12 are instances in which some other voice was taken to be the voice indicated. The number of correct identifications is essentially the same for each of the voices. However, there appears to be a trend toward increasing numbers of false identifications for voices that were rated easy to remember. The correct and false identifications are shown separated by voice and delay in Table I. The significance of the trend toward more false identifications for target voices that were rated easy to remember was tested by the rank-order test Page's L for related samples (Meddis, 1984, pp. 221–224); $L = 41$, $p < 0.05$. The larger number of false identifications engendered by those voices rated easier to remember is the immediate reason that they transmitted less information and were harder signals to detect than were those target voices that were rated hard to remember.

Figure 13 shows the number of times that each of these three voices was correctly or incorrectly identified as the target when it was a probe, i.e., when it was presented at the second listening session. Thus the false identifications shown in Fig. 13 are instances in which the indicated voice was taken to be some other voice. (The correct identifications are the same as in Fig. 12.) In contrast with Fig. 12, we see from Fig. 13 that there is an apparent trend toward more false identifications for the probe voices rated hard to remember. Using Page's L for related samples, the trend was tested on the data separated by delay (shown in bottom columns 3, 5, and 9 of Table II); $L = 42$, $p < 0.025$.

Furthermore, when we examine all the probe voices, we find the same trend towards more false identifications for hard-to-remember probe voices. Figure 14 shows two orders of the voices plotted against each other. On the abscissa, the voices are ordered by their easy-to-remember versus hard-to-remember ratings, and on the ordinate by the number of times each voice was falsely identified as the target. The results were normalized to reflect the lower frequency of the three target voices as potential false identifications. By Spearman's rho, the correlation between the two orderings is 0.77; $p < 0.01$. Thus we see that listeners are more likely to believe erroneously that they have heard hard-to-remember voices previously than to believe erroneously that they have heard easy-to-remember voices previously.

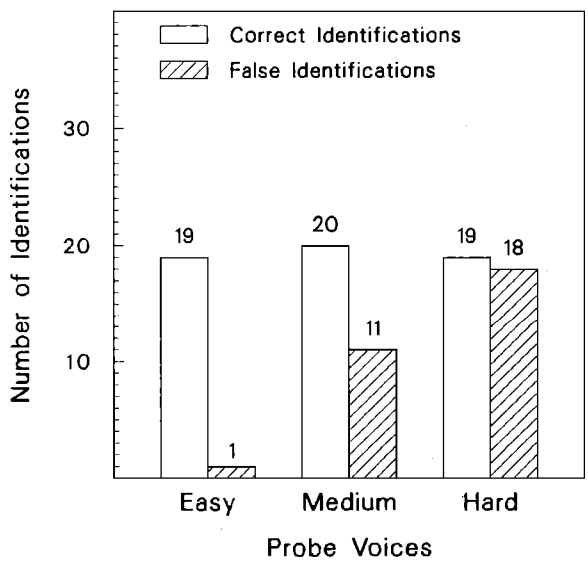

FIG. 13. Number of times that each voice, used as a probe, was correctly or falsely identified as the target voice.

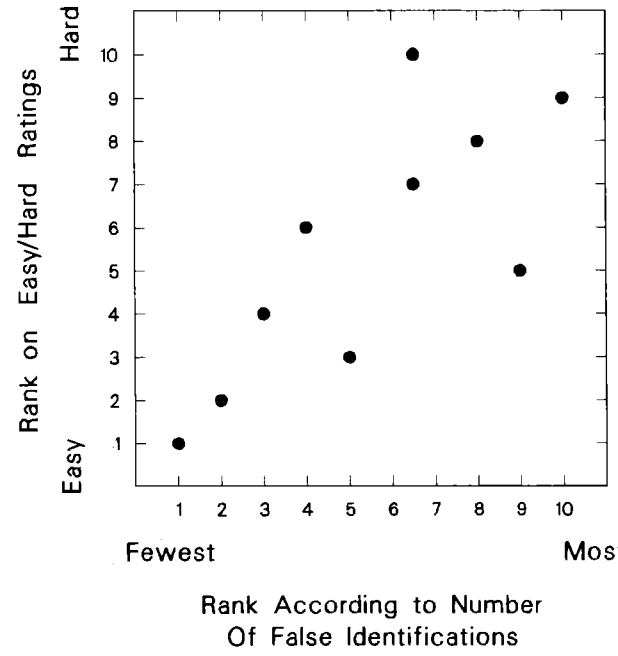

FIG. 14. Rank order of probe voices on easy-to-remember/hard-to-remember ratings versus rank order of the number of false identifications attributable to each probe voice.

Additionally, we calculated the correlation between average certainty that a voice had been heard previously with its "easy/hard" rating: Pearson's $r = 0.72$, $p < 0.05$. In other words, listeners were, on average, more certain that they had heard hard-to-remember probe voices previously.

These results are explicable on the basis of two factors. First, there is a tendency for listeners to maintain a constant rate of correct identifications (Table I, top).

This tendency has been observed previously by Carterette and Barnebey (1975) and by Haggard and Summerfield (1979). In the Carterette and Barnebey experiment, listeners were to say whether a probe voice was among a pool of voices they had heard a short time previously. Carterette and Barnebey observe "In signal detection terms the listener is maintaining a constant hit rate by relaxing his criterion as the signal-to-noise ratio worsens" (pp. 259–260). Haggard and Summerfield used a "same–different" task with varying lengths of reference speech available to the listener. They found that "the different sample conditions tended to produce a fixed percentage of acceptance responses rather than a proportion varying with the available sensitivity in the fashion of an optimal decision maker" (p. 717).

The second factor is the hypothesis that the easy-to-remember voices have relatively unstable representations in memory; hence, their representations decay over time so that they become increasingly similar to, and hence more confusable with, those of the hard-to-remember voices. Thus, if subjects maintain a constant hit rate over all target voices, the number of false identifications will be higher for the easy-to-remember than for the hard-to-remember target voices. This accounts for the result shown in Fig. 12: The features of an easy-to-remember voice that distinguish it readily from other voices are relatively soon lost from memory. As a probe, however, a voice has all its features available to the listener at decision time. Therefore, the probes rated easier to remember will be less frequently mistaken for the target voice than those rated harder to remember. This accounts for the results shown in Figs. 13 and 14.

This tendency for the easy-to-remember voices to be less stable in memory than the hard-to-remember voices is compatible with a theory of prototypes. We now present those aspects of a theory of prototypes that are essential to our account of these data. It has been demonstrated that people use prototypes—central or especially representative members of a category—in classifying members of natural categories such as colors (Berlin and Kay, 1969), animals (Rosch *et al.*, 1975), social and occupational categories (Dahlgren, 1985), and conceptual combinations (Bransford and Franks, 1971; Davis and Papcun, 1987). The principles according to which people use prototypes have been studied experimentally using artificially created categories such as patterns of dots that vary statistically around central values (Posner and Keele, 1968), schematic faces (Reed, 1972; Goldman and Homa, 1977), and geometrical patterns (Franks and Bransford, 1971).

Using somewhat different terminology, Evans and Arnoult (1967, p. 221) explain that "[A] schema is a rule describing a prototype, and a schema family is a population of objects which may be efficiently described in terms of deviations from the prototype." Attneave (1957, p. 81) offered the example that "a figure which may be described as a 'square with a nick on one side' is easier to learn than most other seven-sided polygons because the schema 'square' is simple, familiar, and unambiguous, and the correction, 'with nick in one side' is easily and clearly specifiable." According to this interpretation, nonprototype voices may be effectively characterized in terms of prototypes and deviations therefrom.

Various results indicate that prototypes have a special status in memory. For example, on the basis of observations of serial transmission of folk-stories, diagrammatic and realistic sketches, and other stimuli, Bartlett (1932) suggested that forgetting tends to affect peripheral information more than abstracted prototypical information. With controlled experiments using artificial stimuli, Posner and Keele (1970) found that performance on prototypes decayed more slowly than performance on nonprototypes. In the words of Homa *et al.* (1981): "The stability of prototypical performance, within the context of a deterioration for the old patterns, has now been obtained in numerous experiments (Homa Cross, Cornell, Goldman, and Schwartz, 1973; Homa and Vosburgh, 1976; Strange, Kenney, Kessel, and Jenkins 1970), in which delays varied from 4 days to 10 weeks." Therefore, we should expect that voices which more nearly approach prototypes will be better retained in memory than nonprototypical voices.

According to this line of reasoning, the hard-to-remember voices are prototypes with respect to the other voices in the study. This interpretation is in accord with our findings that probe voices rated hard to remember were more often identified as the target voice than probe voices rated easy to remember, and that listeners were more confident of their heard previously responses the harder to remember a probe voice was rated. Further in accordance with this interpretation, we hypothesize that what listeners remember is a characterization of the voice they heard in terms of a prototype and deviations therefrom. As times passes, listeners lose information about the manner in which the voice they heard deviates from the prototype; in effect, memories of easy-to-remember voices slide toward prototypicality. The process constitutes a psychological analog to statistical regression to the mean.

This interpretation accounts for the overall pattern of asymmetries of errors in the data. When hard-to-remember voices are targets and easy-to-remember voices are probes, there will be relatively few errors because the stable target voice characteristics as well as the immediately present probe voice characteristics are available to the decision. In the converse case, however, an easy-to-remember target voice will lose some of its characteristics. Hence, when a prototypical voice is used as a probe, more errors are to be expected. This trend was tested on the data from the top half of Table II, in which we should expect a drift from few errors at the bottom left to more errors at the upper right. A regression using the row $\times$ column interaction as the independent variable, and the number of errors in each cell as the dependent variable, confirms this hypothesis: $t = 2.71$, with $25df$, $p < 0.012$.

The prototype model of voice recognition also provides a unifying explanation for other previously unconnected findings in the voice recognition literature. For example, it has been found in paired-comparison experiments that the confusability of pairs of voices is not symmetrical (Bricker and Pruzansky, 1966; Dukiewicz, 1970; Thompson, 1985). In forced-choice experiments, such as those cited, a differential tendency to remember prototypical voices over other voices would result in asymmetries such as were observed in those experiments.

Another finding consistent with the prototype model is that listeners take a fundamentally different approach to recognizing familiar and unfamiliar voices, as indicated by studies of brain-damaged and normal subjects, which suggest that injury to either hemisphere impairs the ability to discriminate unfamiliar voices, but only injury to the right hemisphere impairs the ability to recognize familiar voices (Van Lancker *et al.*, 1985; Van Lancker and Kreiman, 1987). It is a natural extension to the prototype model to suggest that, whereas unfamiliar voices are recognized in terms of the prototypes plus deviations, familiar voices are recognized by deviations alone. In other words, when listen-

ers become familiar with a voice, they learn its idiosyncracies and no longer perceive it with respect to a prototype.

Attneave (1957, p. 81) explains how the prototype model can account for the fact that viewers are less accurate in remembering faces or groups of people with whom they are generally unfamiliar: "If the observer has some subjective standard of the human face which he has obtained by 'averaging' the faces of Americans, he may learn a new American face in terms of the manner and degree in which it deviates from this schema (cf. Woodworth's 'correction'). If he is suddenly thrust into a Chinese population, however, his standard will no longer be central, and the new faces will all deviate from it in more or less the same direction." By an analogous argument, the prototype model of memory for voices explains the fact that speakers with dialects unfamiliar to the listener are harder to remember than speakers of familiar dialects (Hollien *et al.*, 1982; Goldstein *et al.*, 1981; Thompson, 1987).

The prototype model makes testable predictions. In particular, in terms of the voices used in the current study, the prototype model predicts that, with repeated exposure to the targets, the listeners would learn to attend to and remember the deviations that characterize each voice without regard to their relationship to a prototype. Then, the relative accuracy for the easy-to-remember voices as compared with the accuracy for the hard-to-remember voices would reverse from the findings of the present study. A more tentative prediction is that, in a study of short-term memory, such as an ABX task, listeners will attend to the idiosyncratic characteristics of each voice; therefore, we would expect greatest accuracy for those pairs in which easy-to-remember voices are compared.

The prototype model also predicts that, at very long delays, listeners will perform at worse than chance levels, because they will select prototypical voices instead of voices they originally heard. We note that, at 5-months' delay, McGehee's listeners achieved only 13% accuracy, whereas chance accuracy would have been 20%. The statistical significance of this result is unclear, however, especially because McGehee took care not to place the targets in positions usually chosen by chance alone.

There are a number of formulations of theories of prototypes and related models (e.g., Bartlett, 1932; Neisser, 1967; Evans, 1967; Wittgenstein, 1953 (see, especially, Secs. 67–77 and 142); Hayes-Roth and Hayes-Roth, 1977; Tversky, 1977; Zadeh, 1965). We do not currently have the means to decide which, if any, among them is most applicable to memory for voices. However, we hope that further experimentation, and especially analysis of similarity ratings data, may help resolve these issues.

Finally, the prototype interpretation is consistent with the predictions of neural network models of learning (see, e.g., McClelland and Rumelhart, 1985). In a neural network model, the prototypes would represent the voices at the bottoms of basins of attraction. As information about nonprototype voices was lost, they would be identified as the voices at the bottoms of the basins of attraction. Thus there is a plausible and realizable computational model that supports the interpretation we have offered in this section.

ACKNOWLEDGMENTS

For help and valuable discussion, we thank Richard Beckman, Jane Booker, Larry Bruckner, Harold Delaney, Ashish Karamchandani, Peter Ladefoged, Tony Warnock, and Jeff Woodard. This work was performed under the auspices of the U.S. Department of Energy under Contract W-7405-Eng.36.

APPENDIX

If each judgment in a sequence of j judgments is independent, the probability of a given sequence of judgments is the product of the probabilities of each of the judgments. There are $\binom{j}{k}$ different sequences of judgments that contain exactly k false identifications. (In this experiment, $j = 9$, inasmuch as one of the voices was the target, and no possibility of false identification existed for that voice.) Hence, assuming independent judgments, if l listeners each make j judgments, the expected number of listeners who make k false identifications and no correct identifications is

$$l\binom{j}{k}[P(S|n)]^k[1-P(S|n)]^{j-k}[P(N|s)],$$

where $P(S|n)$ is the empirically observed probability of false identifications, and $P(N|s)$ is the empirically observed probability of false eliminations.

The expected number of listeners who make k false identifications and one correct identification is

$$l\binom{j}{k}[P(S|n)]^k[l-P(S|n)]^{j-k}[P(S|s)],$$

with the same conventions as above, and where $P(S|s)$ is the observed probability of correct identifications.

[1]Reference to any specific commercial product does not necessarily constitute or imply its endorsement.

[2]We thank Tony Warnock of Cray Research, Inc., for generous assistance in developing this model and running the bootstrap simulation.

[3]Calculating signal-detection measures from group data, as we did, confounds between-subject variation with within-subject variation, leading to lower values of these measures than those calculated from individual subject results (McNicol, 1972, pp. 111–113). However, according to Welford (1986), in cases where the measures have been compared, the differences have been small—about 6%.

[4]Strictly construed, this test requires homogeneity of variance among the delays, which, in fact, does not hold; however, with equal numbers of cases at each delay, the test is robust with respect to violation of this assumption (Box, 1954).

[5]A Pascal program to perform this analysis is available from the first author.

Attneave, F. (**1957**). "Transfer of Experience with a Class-Schema to Identification-Learning of Patterns and Shapes," J. Exp. Psychol. **54**, 81–88.

Attneave, F. (**1959**). *Applications of Information Theory to Psychology: A Summary of Basic Concepts, Methods, and Results* (Holt, New York).

Bartlett, F. C. (**1932**). *Remembering: A Study in Experimental and Social Psychology* (Cambridge U.P., Cambridge, England).

Berlin, B., and Kay, P. (**1969**). *Basic Color Terms: Their Universality and Evolution* (University of California, Berkeley, CA).

Birdsall, T. G., and Peterson, W. W. (**1954**). "Probability of a Correct Decision in a Choice among m Alternatives," University of Michigan: Electronic Defense Group Q. Prog. Rep. No. 10; reprinted in part in *Signal Detection and Recognition by Human Observers: Contemporary Readings*, edited by J. A. Swets (Wiley, New York, 1964); reprinted by Krieger, Malabar, FL.

Box, G. E. P. (**1954**). "Some Theorems on Quadratic Forms Applied in the Study of Analysis of Variance Problems: I. Effect of Inequality of Vari-

ance in the One-Way Classification," Ann. Math. Stat. **25**, 290–302.
Bransford, J. D., and Franks, J. J. (**1971**). "Abstraction of Linguistic Ideas," Cognit. Psychol. **2**, 331–350.
Bricker, P. D., and Pruzansky, S. (**1966**). "Effects of Stimulus Content and Duration on Talker Identification," J. Acoust. Soc. Am. **40**, 1441–1449.
Bricker, P. D., and Pruzansky, S. (**1976**). "Speaker Recognition," in *Contemporary Issues in Experimental Phonetics*, edited by N. Lass (Academic, New York), pp. 295–326.
Carterette, E. C., and Barnebey, A. (**1975**). "Recognition Memory for Voices," in *Structure and Process in Speech Perception*, edited by A. Cohen and S. G. Nooteboom (Springer, New York), pp. 246–265.
Clifford, B. R. (**1980**). "Voice Identification by Human Listeners: On Earwitness Reliability," Law Human Behav. **4**, 373–394.
Clifford, B. R., Rathbone, H., and Bull, H. (**1981**). "The Effects of Delay on Voice Recognition Accuracy," Law Human Behav. **5**, 201–208.
Dahlgren, K. (**1985**). "The Cognitive Structure of Social Categories," Cognit. Sci. **9**, 379–398.
Davis, A. R., and Papcun, G. (**1987**). "The Structure Underlying a Semantic Domain," in *The Mathematics of Language*, edited by A. Manaster-Ramer (Benjamins, The Netherlands).
Diaconis, P., and Efron, B. (**1983**). "Computer-Intensive Methods in Statistics," Sci. Am. **248**(5), 116–130, 170.
Dorfman, D. D., and Alf, E., Jr. (**1969**). "Maximum Likelihood Estimation of Parameters of Signal-Detection Theory and Determination of Confidence Intervals—Rating Method Data," J. Math. Psychol. **6**, 487–496.
Dukiewicz, L. (**1970**). "Frequency-Based Dependence of Speaker Identification," in *Speech Analysis and Synthesis*, edited by W. Jassem (Institute for Fundamental Technical Research, Warsaw), Vol. II, pp. 41–50.
Efron, B. (**1979a**). "Bootstrap Methods: Another Look at the Jackknife," Ann. Stat. **7**, 1–26.
Efron, B. (**1979b**). "Computers and the Theory of Statistics: Thinking the Unthinkable," SIAM Rev. **21**, 460–480.
Evans, S. H. (**1967**). "A Brief Statement of Schema Theory," Psychonom. Sci. **8**, 87–88.
Evans, S. H., and Arnoult, M. D. (**1967**). "Schematic Concept Formation: Demonstration in a Free Sorting Task," Psychononom. Sci. **9**, 221–222.
Franks, J. J., and Bransford, J. D. (**1971**). "Abstraction of Visual Patterns," J. Exp. Psychol. **90**, 65–74.
Goldman, D., and Homa, D. (**1977**). "Integrative and Metric Properties of Abstracted Information as a Function of Category Discriminability, Instance Variability, and Experience," J. Exp. Psychol.: Hum. Learn. Mem. **3**, 375–385.
Goldstein, A., Knight, P., Bailis, K., and Conover, J. (**1981**). "Recognition Memory for Accented and Unaccented Voices," Bull. Psychonom. Soc. **17**, 217–220.
Green, D. M., and Birdsall, T. G. (**1964**). "The Effect of Vocabulary Size on Articulation Score," in *Signal Detection and Recognition by Human Observers: Contemporary Readings*, edited by J. A. Swets (Wiley, New York), pp. 609–619; reprinted by Krieger, Malibar, FL.
Green, D. M., and Swets, J. A. (**1974**). *Signal Detection Theory and Psychophysics* (Krieger, Malabar, FL), 2nd ed.
Haggard, M., and Summerfield, Q. (**1979**) "Perceptual and Memory Factors in Simulated Machine-aided Speaker Verification," Int. J. Man-Mach. Stud. **11**, 717–728.
Hayes-Roth, B., and Hayes-Roth, F. (**1977**). "Concept Learning and the Recognition and Classification of Exemplars," J. Verbal Learn. Verbal Behav. **16**, 321–338.
Hollien, H., Majewski, W., and Doherty, E. (**1982**). "Perceptual Identification of Voices under Normal, Stress, and Disguise Speaking Conditions," J. Phon. **10**, 139–148.
Homa, D., Cross, J., Cornell, D., Goldman, D., and Schwartz, S. (**1973**). "Prototype Abstraction and Classification of New Instances as a Function of Number of Instances Defining the Prototype," J. Exp. Psychol. **101**, 116–122.
Homa, D., Sterling, S., and Trepel, L. (**1981**). "Limitations of Exemplar-Based Generalization and the Abstraction of Categorical Information," J. Exp. Psychol. **7**, 418–439.
Homa, D., and Vosburgh, R. (**1976**). "Category Breadth and the Abstraction of Prototypical Information," J. Exp. Psychol.: Hum. Learn. Mem. **2**, 322–330.
Keppel, G. (**1982**). *Design and Analysis, Second edition* (Prentice Hall, Englewood Cliffs, NJ).
Legge, G. E., Grosmann, C., and Pieper, C. M. (**1984**). "Learning Unfamiliar Voices," J. Exp. Psychol.: Learn. Mem. Cognit. **10**, 298–303.
McClelland, J. L., and Rumelhart, D. E. (**1985**). "Distributed Memory and the Representation of General and Specific Information," J. Exp. Psychol.: General **114**, 159–188.
McEliece, R. J. (**1977**). *The Theory of Information and Coding* (Addison-Wesley, Reading, MA).
McGehee, F. (**1937**). "The Reliability of the Identification of the Human Voice," J. Gen. Psychol. **17**, 249–271.
McGehee, F. (**1944**). "An Experimental Study of Voice Recognition," J. Gen. Psychol. **31**, 53–65.
McNichol, D. (**1972**). *A Primer of Signal Detection Theory* (Allen and Unwin, London).
Meddis, R. (**1984**). *Statistics Using Ranks: A Unified Approach* (Blackwell, Oxford).
Neisser, U. (**1967**). *Cognitive Psychology* (Appleton, Century-Crofts, New York).
Nolan, F. (**1983**). *The Phonetic Bases of Speaker Recognition* (Cambridge U.P., Cambridge, England).
Papcun, G., Ford, W., and Davis, A. R. (**1983**). "Signal Detection Analysis of Voice Input Systems," in Proc. Am. Voice I/O Soc. Voice Data Entry Systems Conference, Chicago, IL, 27–29 Sept. 1983, edited by G. Poock, N. Woo, R. Van Peursem, and M. Joost (Am. Voice I/O Soc., Palo Alto, Cal.).
Posner, M., and Keele, S. (**1968**). "On the Genesis of Abstract Ideas," J. Exp. Psychol. **77**, 353–363.
Posner, M., and Keele, S. (**1970**). "Retention of Abstract Ideas," J. Exp. Psychol. **83**, 304–308.
Reed, S. (**1972**). "Pattern Recognition and Categorization," Cognit. Psychol. **3**, 382–407.
Rosch, E., Mervis, C. B., Gray, W., Johnson, D., and Boyes-Braem, P. (**1975**). "Basic Objects in Natural Categories," Cognit. Psychol. **7**, 573–603.
Rosenberg, A. (**1973**). "Listener Performance in Speaker Verification Tasks," IEEE Trans. Audio Electroacoust. **AU-21**, 221–225.
Saslove, H., and Yarmey, A. D. (**1980**). "Long-Term Auditory Memory: Speaker Identification," J. Appl. Psychol. **65**, 111–116.
Schenker, N. (**1985**). "Qualms about Bootstrap Confidence Intervals," J. Am. Stat. Assoc. Theory Methods **80**, 360–361.
Siegel, S. (**1956**). *Non-parametric Statistics* (McGraw-Hill, New York).
Snedecor, G. W., and Cochran, W. G. (**1967**). *Statistical Methods* (Iowa State U. P., Ames, IA).
Strange, W., Kenney, T., Kessel, F. S., and Jenkins, J. J. (**1970**). "Abstraction over Time of Prototypes from Distortions of Random Dot Patterns: A Replication," J. Exp. Psychol. **83**, 508–510.
Stevens, K. N. (**1972**). "Sources of Inter- and Intra-Speaker Variability in the Acoustic Properties of Speech Sounds," in *Proceedings of the 7th International Congress on Phonetic Science*, edited by A. Rigault and R. Charbonneau (Mouton, The Hague), pp. 206–232.
Swets, J. A., Ed. (**1964**). *Signal Detection and Recognition by Human Observers: Contemporary Readings* (Wiley, New York); reprinted by Krieger, Malibar, FL.
Swets, J. A., and Pickett, R. M. (**1982**). *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory* (Academic, New York).
Thompson, C. P. (**1985**). "Voice Identification: Speaker Identifiability and a Correction of the Record Regarding Sex Effects," Human Learn. **4**, 19–27.
Thompson, C. P. (**1987**). " A Language Effect in Voice Identification," Appl. Cognit. Psychol. **25**, 121–131.
Tippett, L. H. C. (**1925**). "On the Extreme Individuals and the Range of Samples Taken from a Normal Population," Biometrika **17**, 364–387.
Tversky, A. (**1977**). "Features of Similarity," Psychol. Rev. **84**, 327–352.
Van Lancker, D., and Kreiman, J. (**1987**). "Voice Discrimination and Recognition Are Separable Abilities," Neuropsychologia **25**, 829–834.
Van Lancker, D., Kreiman, J., and Cummings, J. (**1985**). "Voice Recognition and Discrimination: New Evidence for a Double Dissociation," J. Clin. Exp. Neuropsychol. **7**, 609.
Welford, A. T. (**1986**). "Two Comparisons of Recognition and Recall by Signal Detection Measures," Br. J. Psychol. **77**, 237–242.
Wittgenstein, L. (**1953**). *Philosophical Investigations* (Blackwell, Oxford); translated by G. E. M. Anscombe.
Woodworth, R. S. (**1938**). *Experimental Psychology* (Holt, New York).
Zadeh, L. A. (**1965**). "Fuzzy Sets," Inf. Control **8**, 338–353.