

Comparing discrimination and recognition of unfamiliar voices

Jody Kreiman

Division of Head and Neck Surgery, UCLA School of Medicine, and Audiology and Speech Pathology (126), West Los Angeles VA Medical Center, Wilshire and Sawtelle Blvds., Los Angeles, CA 90073, USA

George Papcun

Los Alamos National Laboratory, USA

Received 16 October 1990

Revised 11 March 1991

Abstract. This study compared listeners' abilities to discriminate among and remember the voices of 10 young male Californians, and examined the strategies listeners used in the two tasks. One group of listeners ($n = 24$) judged whether pairs of voices represented the same or two different speakers; a second ($n = 100$) heard a single voice and then tried to identify it in a "voice lineup" one week later. Signal detection analyses revealed no significant differences between discrimination and recognition accuracy. Both correct and incorrect "same" responses were more frequent for the voice discrimination task than for the recognition task. Multidimensional scaling analyses suggested that forgetting does not affect all voice "features" equally: some kinds of information were lost, while others were well-preserved over time. Comparisons of patterns of confusions for the two tasks supported the notion that voices are remembered in terms of a "prototype" and a set of deviations from that prototype, and that over time the deviations are forgotten so that identification responses converge on the most "typical" sounding voices.

Zusammenfassung. Diese Studie vergleicht die Fähigkeit von Hörern zwischen den Stimmen von zehn jungen kalifornischen Sprechern zu unterscheiden und sie wiederzuerkennen. Die Studie untersucht auch die Strategien welche von den Hörern während den zwei Aufgaben angewandt wurden. Eine Gruppe von Hörern ($n = 24$) entschied ob ein Paar von Stimmen von dem gleichen Sprecher oder von zwei verschiedenen Sprechern stammten; eine zweite Gruppe ($n = 100$) hörte eine Stimme und versuchte sie wiederzuerkennen anhand einer "Stimmen-gegenüberstellung" welche eine Woche später stattfand. Die Analyse zeigte keinen Unterschied auf zwischen der Genauigkeit der Erkennung und der der Unterscheidung. Sowohl korrekte als auch unkorrekte "identische" Antworten waren zahlreicher im Falle der Aufgabe der Unterscheidung. Eine Analyse anhand einer mehrdimensionalen Skalierung weist darauf hin, daß nicht alle Stimmerkmale gleich schnell vergessen werden. Ein Vergleich der Muster der Verwechslungen während den zwei Aufgaben suggeriert, daß Stimmen anhand eines Prototyps und den Abweichungen von demselben gespeichert werden. Im Laufe der Zeit werden die Abweichungen vergessen und die Identifizierung tendiert auf die meist typischen Stimmen zu.

Résumé. Cette étude compare la faculté d'auditeurs de se rappeler et de discriminer entre les voix de dix jeunes Californiens et elle examine les stratégies de ces auditeurs pendant les deux tâches. Un premier groupe d'auditeurs ($n = 24$) jugeait si des couples de voix provenaient ou non d'un même locuteur; un deuxième groupe ($n = 100$) écoutait une voix unique et tentait alors de l'identifier sur la base d'une "confrontation vocale" qui avait lieu une semaine après. Des méthodes de détection du signal ne montrent pas de différence entre la précision de la discrimination et la précision de la reconnaissance. Les réponses "identiques" exactes et inexactes étaient plus fréquentes lors de la tâche de discrimination que lors de la tâche de reconnaissance. Une analyse par échelle multidimensionnelle suggère que l'oubli n'affecte pas de manière égale tous les indices de la qualité vocale: certains types d'information se perdaient, d'autres étaient préservés dans le temps. Une comparaison des formes de confusions pour les deux tâches suggère l'idée que les voix sont mémorisées en termes d'un prototype et d'un ensemble de déviations par rapport à celui-ci. Au cours du temps, les déviations seraient oubliées de telle sorte que les réponses d'identification convergeraient vers les voix les plus typiques.

Keywords. Speaker recognition, voice discrimination, prototype theory.

1. Introduction

Relatively little is known about how listeners perceive and remember unfamiliar voices, despite a long history of research on voice quality in linguistics, otolaryngology, speech science, psychology and other disciplines. Understanding how stored representations of voices change over time has obvious applications, including designing efficient automatic voice recognition systems. However, no study has directly compared voice discrimination and recognition, although several studies have examined listeners' abilities to recognize unfamiliar voices after short and relatively long delays. For example, Clifford (1980; Clifford et al., 1981) compared recognition scores for voices previously identified as easy- or hard-to-recognize. Subjects attempted to recognize these voices after delays of 10, 40, 100 or 130 minutes in an 11-alternative forced choice task. Clifford found that the percentage of correct responses declined over time for hard-to-recognize voices, but not for easily-recognized voices. A second experiment compared delays of 10 minutes, 24 hours, 7 days and 14 days; only the shortest delay differed significantly from the others. Saslove and Yarmey (1980) examined the effect of three factors – speaker's tone of voice (angry versus calm), anticipated versus surprise testing, and "immediate" testing (a few minutes after exposure) versus a 24 hour delay between learning and testing – on voice recognition scores, using a five-alternative forced-choice task. Both hit rates (correct "heard previously" responses) and false alarm rates (incorrect "heard previously" responses) varied significantly with tone-of-voice condition and with instructions: hit rates were higher, and false alarm rates were lower, when subjects knew a test was coming and when the tone of the test voice matched that of the original stimulus. No effect of delay was observed. Legge et al. (1984) varied the size of the retention set as well as delay; listeners heard five to twenty voices, which they tried to identify after fifteen minutes or ten days in a series of two-alternative forced-choice trials. Again, delay did not significantly affect recognition scores, although scores improved as the size of the retention set decreased.

The present study compared voice discrimination and recognition accuracy directly, and also examined differences in the strategy used to recognize unfamiliar voices after one week, as compared to that used to discriminate among the same voices in a paired comparison task. The goals were (1) to determine how hit rates, false alarm rates and overall accuracy vary with task demands for a single set of voices; and (2) to determine whether some kinds of information about voice quality are better retained than others in memory. To our knowledge, no evidence currently exists as to what listeners remember about voices over time.

2. Method

2.1. Stimuli

Selection of the stimulus voices has been described in detail elsewhere (Papcun et al., 1989). Briefly, 22 male Southern Californians (all native speakers of English) were recorded making telephone survey calls. This format controlled the content of the voice samples while allowing speakers to talk in a natural, interactive context. Each speaker made two recordings at each of two sessions separated by at least one week. Recordings were made in a quiet office on a Uher 4200 reel-to-reel tape recorder via a high-quality microphone attached to the mouth-piece of the telephone. One call was selected from each session, based on naturalness, fluency, conformity to the script, and lack of extraneous comments. Tapes were edited to remove comments not in the script and to shorten lengthy pauses. Edited calls lasted an average of 1.58 minutes (sd = 0.13 minutes).

Seven speakers who strayed too far from the survey text were eliminated from the voice set. The remaining 15 voices were played to 50 listeners who judged, for each voice, how easy or hard they felt it would be to remember (on a seven-point scale). Ratings for each voice were summed across listeners and standardized. Two voices that were described as accented by an appreciable number of raters were eliminated from the response set. Based on a cumulative semi-normal plot of standardized scores versus the rank order

of the remaining voices (from easiest to hardest to remember), the ten speakers that best approximated a normal distribution of distinctiveness¹ (from easy- to hard-to-remember) according to the listeners' ratings were selected for use in these experiments. Speakers were otherwise unselected with respect to voice quality.

The final ten speakers ranged in age from 19 to 31 years. All were free of vocal pathology and accent other than that of Southern California.

2.2. Listeners

One hundred native speakers of English, ranging in age from 17–47 (mean age 25.94 years, sd 7.14 years), participated in the voice recognition portion of this study. Twenty-four native speakers of English, ranging in age from 21–45 (mean age 29.63 years, sd 6.08 years), took part in the voice discrimination study. All listeners reported normal hearing. No listener participated in both studies.

2.3. Voice discrimination task

Three sentences were excerpted from each of the two survey calls, for a total of six stimulus tokens/speaker. Each sentence had at least eight syllables and lasted approximately two seconds. Stimuli were edited on the computer waveform editor at the UCLA Phonetics Laboratory. Tapes were played at half speed (3 3/4 ips); they were low-pass filtered at 4 kHz and sampled at 10 kHz, for an effective filter setting of 8 kHz and an effective sample rate of 20 kHz.

A stimulus tape was constructed using both orders (AB and BA) of all possible pairs of the ten voices (90 "voices different" trials), plus an additional 30 pairs (three stimulus sentences \times 10 voices) where voices were the same, for a total of 120 pairs of voices. Different sentences were used for the AB and BA orders of each pair, but within

a pair of voices speakers always said the same thing. For "voices different" trials, utterances were both taken from calls made at the same recording session (first or second); for "voices same" trials, one utterance was taken from each of the two calls, so listeners never compared two identical stimulus tokens. Note that the difficulty of "voices same" trials depended on the amount of sampled within-speaker variability in voice quality, and the difficulty of the "voices different" trials depended on sampled between-speaker differences. The task for listeners thus consisted of distinguishing these two sources of variation, i.e., in deciding whether differences between samples were due to variations in the vocal quality of a single speaker, or to differences between two speakers.

Each utterance type occurred an equal number of times. Which token of an utterance (i.e., from the first or second survey call) was used in a given pair was determined by random assignment, with the constraint that each token appeared an equal number of times. For "voices same" trials, the version of a sentence that occurred first in a pair was varied at random, with each occurring first an equal number of times. Voices within a pair were separated by 1.5 seconds, and pairs were separated by 12 seconds.

Three additional pairs of voices were constructed using voices not in the stimulus set. These pairs were used as practice to familiarize listeners with the test procedures and response scales (described further on).

Listeners were tested in groups of no more than five. Stimuli were presented to all listeners in the same random order. The stimulus tape was played over a loudspeaker at a constant comfortable listening level. Testing took place in a quiet room; listeners were seated no more than three feet from the loudspeaker. Listeners were instructed to judge each pair of voices as independently as possible, to determine whether voices within a pair were the same or different, and to rate their confidence in each same/different response and the dissimilarity of the voice pairs (both on five-point scales).

Before the actual listening test, listeners heard the three practice pairs of voices and were given feedback as to the correctness of their responses.

¹ In the absence of evidence to the contrary, we assume that, across a population of speakers, most voices are neither especially easy nor especially hard to remember, but instead are somewhere in between. The precise form of the true population distribution, however, is of little importance here.

Breaks were given after the fortieth and eightieth trials. The entire session lasted about 40 minutes.

2.4. Voice recognition task

Methods for this task are described in detail in Papcun et al. (1989). Listeners were assigned at random to one of ten groups. At a first listening session, each group heard a single target voice; the full survey call from the first recording session was played. Listeners were told they were about to hear the voice of a young male Californian, and were asked to pay close attention to the voice because they would have to identify it later.

After exactly one week listeners returned and heard a series of voices. They were told the target voice might or might not be presented or might be presented more than once, although in reality each of the ten voices (including the target) was presented once only. Full survey calls from the second recording session were played at this listening session, so listeners again were unable to use a simple stimulus-matching strategy. For each voice, listeners reported sameness/difference, their confidence in their response, and the similarity of the stimulus to the remembered target, using five-point scales as above.

2.5. Multidimensional scaling analyses

For the voice discrimination task, similarity ratings for individual listeners were assembled into matrices. These were examined for consistent

patterns of asymmetry; since none were found, matrices were symmetrized by averaging across the diagonal and analyzed using three-way (individual differences) nonmetric multidimensional scaling (the SAS procedure ALSCAL (SAS Institute, 1983)). Solutions were found in 2–6 dimensions. Stress and r^2 values for each solution are given in Table 1. Stress measures how far the data depart from the multidimensional scaling model, and r^2 measures the amount of variance in the underlying data accounted for by the scaling solution (see e.g. Schiffman et al., 1981). Based on these values and on interpretability, the four-dimensional solution was selected ($r^2 = 0.71$).

For the voice recognition task, similarity and confidence ratings for each target voice group were also assembled into full matrices. Each cell in those matrices contained the average of the similarity or confidence ratings made by the ten listeners in that target voice group. Since substantial asymmetries occurred, these matrices were not averaged across the diagonal, but were treated as asymmetrical by the scaling program. Further, data were treated as row-conditional, since columns and rows have different statuses in these matrices (columns representing the voices are remembered targets, and rows representing the voices as newly-presented foils). Three-way non-metric scaling solutions were found in 2–6 dimensions. Stress and r^2 values are included in Table 1. The three-dimensional solution was selected for further analysis ($r^2 = 0.91$).

2.6. Data recoding procedure

For both the discrimination and recognition tasks, confidence ratings were combined with same/different responses to produce a single 10-point scale ranging from “positive voices are the same” (1) to “positive voices are different” (10). A “same” response combined with a confidence rating of 1 (“positive response is correct”) was coded as 1 in the new scale; a “same” response combined with a confidence rating of 5 (“very uncertain response is correct”) was coded as 5; a “different” response combined with a confidence rating of 5 was coded as 6; a “different” response combined with a confidence rating of 1 was coded as 10 on the new scale; and so on for the interven-

Table 1
Stress and r^2 values for the multidimensional scaling solutions

Task	# Dimensions	Stress	r^2
Discrimination	2	0.235	0.606
	3	0.165	0.669
	→ ^a 4	0.121	0.709
	5	0.093	0.717
	6	0.067	0.750
Recognition	2	0.195	0.838
	→ ^a 3	0.124	0.910
	4	0.086	0.946
	5	0.061	0.964
	6	0.052	0.977

^a Arrows mark the solution selected.

ing scale values. These unfolded ratings were used in the signal detection analyses described in Section 3.1.

2.7. Measured and rated characteristics of the voices

The multidimensional scaling solutions were interpreted by examining the correlations between each voice's coordinates on the derived dimensions and a set of ratings and measurements of that voice. The acoustic and perceptual characteristics used are listed in Table 2. Wide and narrow band spectrograms were produced for all 60 stimulus utterances. Mean fundamental frequency (F_0) was determined by measuring the frequency of the tenth harmonic at the midpoint of each vowel and averaging these values across all vowels and utterances for a given speaker. F_0 standard deviation is the standard deviation associated with each mean value; F_0 range is the difference between the highest and lowest F_0

value observed for each speaker. The duration of each utterance, determined by measuring from the onset of voicing to the offset of voicing on wideband spectrograms, was divided by the number of (phonetic) syllables in that utterance. The average of these values across utterances is each speaker's mean speaking rate, and their standard deviation is the rate standard deviation. Values of F_1 , F_2 and F_3 were measured at the steadiest portion of each vowel and then averaged across vowels and utterances for each speaker; the difference between F_2 and F_1 ($F_2 - F_1$) and the ratio of F_2 to F_1 (F_2/F_1) were calculated using these mean values.

Judgments of the perceptual characteristics of the stimulus voices were obtained by having five phoneticians rate each voice on the 17 scales included in Table 2. These scales were drawn in part from previous multidimensional scaling and factor analytic studies of voice quality (e.g., (Voiers, 1964; Holmgren, 1967; Matsumoto et al., 1973; Carterette and Barnebey, 1975; Murry et al., 1977; Singh and Murry, 1978; Walden et al., 1978; Murry and Singh, 1980; Fagel et al., 1983; Kempster, 1984)), and in part from the literature on personality (e.g., (Comrey, 1973)). All raters were native speakers of American English, and all had experience evaluating vocal quality. Raters were provided with a tape on which each speaker's six utterances were recorded in sequence, and with a set of rating sheets listing the 17 scales. They were instructed to listen to each voice as often as they liked, and to rate it on each of the scales by making a mark at the appropriate point on a line that connected the two endpoints of the scales. Raters were instructed to consider the endpoints of each scale to represent, in their best judgment, extreme values on that attribute for normal college-aged male Southern Californians. The highest and lowest ratings for each voice/scale pair were discarded, and the remaining three ratings were summed to produce a single score for each voice on each scale.

Table 2
Measured and rated acoustic and perceptual characteristics of the stimulus voices

Measured characteristics	
Mean fundamental frequency (F_0) (Hz)	
F_0 standard deviation	
F_0 range (Hz)	
Mean speaking rate (syllables/second)	
Rate standard deviation	
Mean frequency of F_1 (Hz)	
Mean frequency of F_2 (Hz)	
Mean frequency of F_3 (Hz)	
$F_2 - F_1$	
F_2/F_1	
Rated characteristics	
Active/lacking energy	Light voice/weighty voice
Bored/involved	Loud voice/soft voice
Breathy voice/clear voice	Masculine/feminine
Casual/formal	Monotone voice/variable pitch
Creaky voice/smooth voice	Pleasant voice/unpleasant voice
Extraverted/introverted	Sincere/insincere
Fast talker/slow talker	Tense, nervous/relaxed
Fluent/dysfluent	
Happy/sad	
High-pitched voice/low-pitched voice	

3. Results

3.1. Accuracy of discrimination versus accuracy of recognition

Receiver operating characteristics (ROCs) were constructed using the unfolded voice ratings described in Section 2.6. These curves are presented in Figure 1; they show the cumulative probability of a correct "voices same" response (the hit rate) on the ordinate, plotted against the cumulative probability of an incorrect "voices same" response (the false alarm rate) on the abscissa. Least-squares estimates of A_z (the area under the binormal ROC) and d'_z (another signal detection theoretic measure of accuracy) were calculated using a program by Dorfman and Alf (1969; reprinted in (Swets and Pickett, 1982)). Results are given in Table 3, along with hit rates and false alarm rates for the two tasks. A_z values did not differ significantly for the two tasks ($z = 1.016$, n.s. (Swets and Pickett, 1982, pp. 80–88)): the ability of listeners to discriminate among these voices was not significantly greater than their ability to recognize a single voice after one week.

Listeners did differ in their response biases for the two tasks, however. As Table 3 shows, the hit rate for the voice discrimination task was 25% higher than that for the voice recognition task. Correspondingly, subjects produced nearly 14%

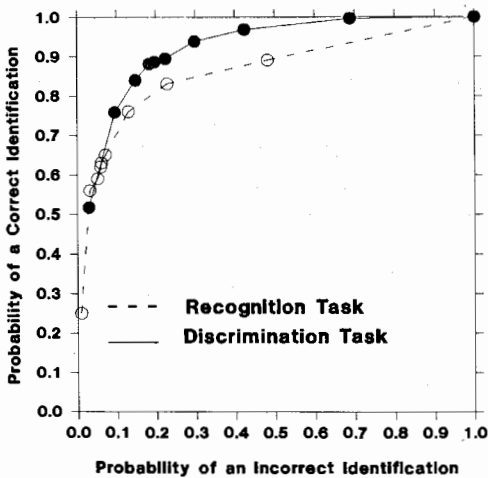


Fig. 1. Receiver operating characteristics (ROCs) for the voice discrimination and recognition tasks.

Table 3

Accuracy measures for the voice discrimination and voice recognition tasks

Measure	Task	
	Discrimination	Recognition
A_z	0.923 (0.005)	0.873 (0.021)
d'_z	2.022 (0.054)	1.625 (0.142)
Hit rate	88.6% (638/720)	63% (63/100)
False alarm rate	19.7% (426/2160)	6.11% (55/900)

more false alarms in the discrimination task than did subjects in the recognition task. An examination of the ROCs in Figure 1 suggests that this relative bias toward "same" responses in the discrimination task held for all levels of confidence: points on the curve for the discrimination task are all higher and to the right of corresponding points on the curve for the recognition task, indicating that both hit and false alarm rates were greater for each level of confidence.

3.2. Patterns of confusions

We first examined the number of times a given voice was falsely identified as some other target, i.e., how often each voice received an incorrect "same" response when it was a foil (see Figure 2(a)). (For the voice discrimination task, the first voice in a pair was considered the "target".) For the recognition task, a voice rated harder-to-remember was significantly more likely to be mistaken for another target than was an easy-to-re-

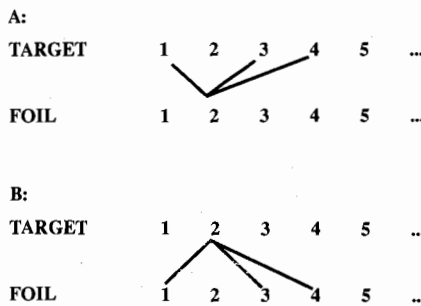


Fig. 2. Patterns of confusions for the two tasks. (a) The number of times a given voice was falsely identified as some other target. (b) The number of times some other voice was incorrectly identified as a given target.

member voice (Spearman's $\rho = 0.80$, $p < 0.01$). This confirms our previously-reported finding from an experiment using three target voices (Papcun et al., 1989). In contrast, there is no significant relationship between easy/hard-to-remember ratings of foils and the frequency of mistaken identification for the voice discrimination data (Spearman's $\rho = 0.46$, n.s.). Complete confusion matrices for both tasks are included as an appendix to this report.

We next examined the number of times some other voice was incorrectly identified as a given target (Figure 2(b)). For the discrimination task, fewer false alarms occurred when the target was rated easy-to-remember than when it was rated hard-to-remember (Spearman's $\rho = 0.65$, $p < 0.05$). For the recognition task, the opposite pattern occurred: fewer false alarms were observed when the target was rated *harder* to remember than when it was easy-to-remember (Spearman's $\rho = -0.67$, $p < 0.025$). The frequency with which a voice was *correctly* identified when it was a target (the hit rate) was not significantly related to the easy-to-remember/hard-to-remember ratings for either the discrimination or recognition tasks.

3.3. Multidimensional scaling results

Similarity ratings for the discrimination and recognition tasks were significantly, but not highly, correlated (Pearson's $r = 0.64$, $p < 0.01$). An examination of the multidimensional scaling solutions for the two tasks (derived from these similarity ratings) revealed several changes in patterns of similarity with time.

The three-dimensional solution for the voice recognition data accounts for 91% of the variance in confidence and similarity ratings, as compared to 71% for the four-dimensional solution selected for the voice discrimination data. The first dimensions (*D1*) in the two spaces are significantly correlated ($r = -0.88$, $p < 0.01$), and thus apparently represent roughly the same perceptual "feature". Table 4 lists the significant correlations between stimulus coordinates on these dimensions and measured and rated characteristics of the voices. R^2 values in this table reflect the amount of variance in the underlying data which is ac-

Table 4

Significant correlations between the first dimensions and characteristics of the voices (all correlations are significant at $p < 0.05$)

Voice discrimination space	Voice recognition space
$r^2 = 0.35$	$r^2 = 0.45$
Pleasant/unpleasant ($r = 0.92$)	Pleasant/unpleasant ($r = -0.74$)
High/low pitched ($r = -0.84$)	High/low pitched ($r = 0.80$)
Tense/relaxed ($r = -0.78$)	Tense/relaxed ($r = 0.58$)
Light/heavy ($r = -0.75$)	Light/heavy ($r = 0.75$)
$F2$ ($r = 0.73$)	$F2$ ($r = -0.77$)
Mean $F0$ ($r = 0.68$)	Mean $F0$ ($r = -0.74$)
$F2 - F1$ ($r = 0.67$)	$F2 - F1$ ($r = -0.76$)
Masculine/feminine ($r = 0.66$)	Masculine/feminine ($r = -0.66$)
$F3$ ($r = 0.64$)	$F3$ ($r = -0.63$)
	$F0$ range ($r = -0.62$)
	$F2/F1$ ($r = -0.61$)
	$F0$ SD ($r = -0.60$)
	Bored/interested ($r = -0.56$)

Table 5

Correlations between dimensions in the perceptual spaces for the discrimination and recognition tasks

Discrimination space	Recognition space		
	Dimension 1	Dimension 2	Dimension 3
Dimension 1	-0.88 ^a	-0.37	-0.36
Dimension 2	0.02	0.36	0.27
Dimension 3	-0.28	0.06	0.09
Dimension 4	-0.18	-0.18	0.58

^a $p < 0.05$ (adjusted for multiple comparisons).

counted for by the dimension. Note the increased importance of *D1* in the perceptual space for the recognition task ($r^2 = 0.45$ versus 0.35).

No other significant correlations were observed between dimensions in the discrimination and recognition spaces (Table 5). Acoustic and perceptual correlates of the remaining dimensions in both spaces are listed in Table 6.

4. Discussion

Multidimensional scaling analyses of data from the voice discrimination task revealed four perceptual features (which may be labelled "mascu-

Table 6
Correlations between remaining dimensions and voice characteristics (all correlations are significant at $p < 0.05$)

Task	Dimension	r^2	Correlates
Discrimination	2	0.158	Creaky/smooth ($r = 0.70$)
	3	0.114	Loud/soft ($r = 0.71$) Light/heavy ($r = -0.66$) F0 range ($r = 0.65$) Masculine/feminine ($r = 0.63$)
	4	0.087	Happy/sad ($r = -0.77$)
Recognition	2	0.247	Breathy/clear ($r = -0.61$)
	3	0.213	Rate SD ($r = 0.69$) Happy/sad ($r = -0.63$) Creaky/smooth ($r = 0.62$) Tense/relaxed ($r = 0.61$) Monotone/variable ($r = 0.60$) Fluent/dysfluent ($r = -0.56$) Pleasant/unpleasant voice ($r = -0.56$) Bored/interested ($r = 0.55$)

linity”, “creakiness”, “variability” and “mood”), which together accounted for 71% of the variance in the underlying data. Confidence and similarity ratings from the voice recognition task were even better predicted by only three perceptual dimensions (“masculinity”, “breathiness” and “liveliness”). Each of these dimensions accounted for a large amount of the variance in the underlying data (45.6%, 24.7% and 21.3%, respectively). The “features” revealed by these multidimensional scaling analyses properly describe only the present voice set, and are not necessarily generalizable to other sets of speakers. However, comparison of r^2 values and of the derived dimensions for the two solutions (derived from the same set of stimuli) suggest that listeners do not simply remember the same things they listen for in discrimination tasks, but imperfectly: some things are lost, and others correspondingly gain in importance, in memory over time.

Changes in stored representations with time do not imply a loss of ability to distinguish among voices. Recognition accuracy did not differ significantly from discrimination accuracy for the voices and delays studied here, although both hit and false alarm rates were elevated for the discrimination task relative to the recognition task. It is possible that this finding reflects the difference in stimulus durations in the two tasks: sub-

jects in the discrimination task heard pairs of sentences, while listeners in the recognition task heard entire survey calls. However, voice discrimination performance has been shown to peak with stimulus durations of 500–1000 ms (Pollack et al., 1954; Bricker and Pruzansky, 1966), suggesting increasing stimulus durations would not have increased our subjects’ discrimination performance beyond observed levels.

The discrimination and recognition tasks did apparently require subjects to adopt rather different strategies. Differences in task demands may have biased listeners overall toward more or less stringent response criteria. Recall that listeners never compared two identical stimuli in either task; thus subtle differences in pronunciation or intonation contours (for example) might distinguish two speakers with similar voices, or two different productions by the same speaker. Listeners in the voice discrimination task were therefore forced to accept some amount of difference among stimuli as being consistent with a “same” response, and consequently adopted a fairly loose response criterion. In contrast, listeners in the voice recognition task may have adopted a relatively strict response criterion as a result of our “open set” design. These listeners were told the target voice might or might not occur in the recognition set, which may have discouraged them

from responding "same" too readily.

As response criteria and kinds of features used varied across tasks, so did patterns of confusions. The probability of *correctly* identifying a target voice (the hit rate) was not related to the target's rated distinctiveness, for either the discrimination or recognition tasks. However, which voices were confused did depend on both the task and the distinctiveness of the voices. As might be expected, for the discrimination task targets (the first members of voice pairs) rated "easy to remember" were significantly less likely to be *falsely* identified as some other voice than were targets rated "hard to remember", but the distinctiveness of the foils (i.e., second members of voice pairs) was not significantly related to false alarm rates. This suggests that the first member of a pair provides "context" for the judgment: the two voices are not equal in this task. For the voice recognition task, target voices rated *hard* to remember were less likely to be associated with incorrect "same" responses, while foils rated hard to remember were *more* likely to be incorrectly identified as some other target. These results are summarized in Table 7.

These findings are consistent with the claim that listeners remember voices in terms of a "prototype" – a typical voice or central category member – and a set of deviations from that prototype (Papcun et al., 1989). Papcun et al. argued that "hard to remember," or average-sounding, voices are prototypes with respect to the other voices studied here. Information about how voices deviate from a prototype is apparently rather unstable, and is forgotten relatively quickly. Thus Papcun et al. found that, over time,

listeners increasingly forgot how a target voice differed from the prototype, and their memories converged on the hard-to-remember voices no matter what voice they originally heard.

This model accounts for differences in patterns of confusions in the present study. For the discrimination task, where only 1.5 seconds intervened between hearing a voice and "identifying" it, most information about the target voice is still available when a "same" or "different" decision must be made. Thus easy-to-remember target voices attract fewer false alarms, and the distinctiveness of foil voices is not related to false alarm rates. However, the tendency to select a typical-sounding voice increases as information is lost about how the target voice differed from the prototype. As a result, the false identification rate associated with harder-to-remember targets decreases with time: since they are average-sounding, the drift toward a prototype in memory generates *correct* responses for these voices. Correspondingly, more distinctive (that is, non-prototypical) targets are associated with more false identifications, because the average-sounding voices are mistakenly selected in their stead.

Assuming that there is also a tendency for less-distinctive targets to be remembered less well than more distinctive ones, this model also explains the lack of change in the hit rate with voice distinctiveness: the tendency to select average-sounding voices favors these targets, while a tendency to better remember distinctive voices favors those targets. These two hypothetical sources of correct "same" responses for average-sounding voices (i.e., actual correct identifications versus correct identifications due to drift toward a prototype) unfortunately cannot be separated in the present data.

It is not entirely clear how to combine information from the multidimensional scaling results with a prototype model, because the multidimensional scaling spaces were derived from similarity ratings and the prototype model from patterns of confusions among voices. Confusions are not necessarily predictable from similarity ratings: for example, two very similar voices may never be confused, or may always be confused. In the present study, the first dimension ("masculinity") appeared in the multidimensional scaling solutions

Table 7
Patterns of false alarms for the discrimination and recognition tasks

Task	Target	Foil
Discrimination	Fewer false alarms when easy-to-remember	No difference in false alarm rates with distinctiveness
Recognition	Fewer false alarms when hard-to-remember	More false alarms when hard-to-remember

for both the discrimination and recognition data, suggesting that this "feature" is somehow central for the characterization of this population of speakers. However, further investigations using many more stimulus voices will be necessary to determine precisely what information a prototype includes and what role (if any) it plays in similarity judgments.

Acknowledgments

Thanks are due to Peter Ladefoged, Ian Maddieson, and the UCLA Phonetics Lab Group for much helpful discussion and for use of their facilities. Thanks also to Kristin Precoda for adapting the signal detection analysis program for use on our hardware. This research was supported in part by an NINCDS post-doctoral traineeship to the first author (NS07059).

Appendix A. Confusion matrices

Columns represent the voices as targets (or as the first member of a pair for the discrimination task); rows represent the same voices as foils (or as the second member of a pair). The diagonal thus represents the number of correct "same" responses for each task, and the off-diagonal entries represent incorrect "same" responses. Off-diagonal entries are summed in the margins.

A.1. Voice discrimination task (maximum n/cell = 24). Entries on the diagonal represent the average of three presentations

Voice rank order (Easy-to-remember-hard-to-remember)											
	1	2	3	4	5	6	7	8	9	10	
1	23	3	0	2	19	1	1	9	10	2	47
2	4	22	1	2	1	0	2	10	4	10	34
3	1	0	19	11	1	9	7	0	3	1	33
4	0	0	10	19	4	5	5	2	2	8	36
5	1	3	10	2	18	6	5	2	7	2	38
6	2	0	7	6	11	23	8	6	1	7	48
7	1	3	10	5	12	13	21	1	1	7	53
8	9	9	2	0	6	0	0	23	8	1	35
9	2	8	5	1	13	1	10	8	22	14	62
10	3	10	0	0	5	1	3	7	11	22	40
	23	36	45	29	72	36	41	45	47	52	

A.2. Voice recognition task (maximum n/cell = 10)

Voice rank order (Easy-to-remember-hard-to-remember)										
	1	2	3	4	5	6	7	8	9	10
1	5	0	0	0	0	0	0	0	0	0
2	0	6	0	0	0	0	0	0	0	0
3	0	0	7	1	1	0	2	0	0	4
4	0	0	0	5	0	2	1	0	1	0
5	2	0	2	1	7	0	0	0	1	0
6	2	0	0	0	0	4	1	0	0	3
7	1	1	1	0	1	1	9	1	2	1
8	0	3	0	0	1	0	1	6	0	5
9	2	2	1	1	2	2	4	2	7	18
10	2	1	1	1	0	0	0	1	0	7
	9	7	5	4	5	5	9	4	4	3

References

P.D. Bricker and S. Pruzansky (1966), "Effects of stimulus content and duration on talker identification", *J. Acoust. Soc. Amer.*, Vol. 40, pp. 1441-1449.

E.C. Carterette and A. Barnebey (1975), "Recognition memory for voices", in *Structure and Process in Speech Perception*, ed. by A. Cohen and S. Nooteboom (Springer, New York), pp. 246-265.

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

B. Clifford, H. Rathbone and H. Bull (1981), "The effects of delay on voice recognition accuracy", *Law and Human Behavior*, Vol. 5, pp. 201-208.

A.L. Comrey (1973), *A First Course in Factor Analysis* (Academic Press, New York).

D.D. Dorfman and E. Alf, Jr. (1969), "Maximum likelihood estimation of parameters of signal-detection theory and determination of confidence intervals-rating method data", *J. Math. Psych.*, Vol. 6, pp. 487-496.

W.P.F. Fagel, L.W.A. van Herpt and L. Boves (1983), "Analysis of the perceptual qualities of Dutch speakers' voice and pronunciation", *Speech Communication*, Vol. 2, No. 4, pp. 315-326.

G.L. Holmgren (1967), "Physical and psychological correlates of speaker recognition", *J. Speech Hearing Res.*, Vol. 10, pp. 57-66.

G. Kempster (1984), *A multidimensional analysis of vocal quality in two dysphonic groups* (PhD Dissertation, Northwestern University).

G.E. Legge, C. Grosman and C.M. Pieper (1984), "Learning unfamiliar voices", *J. Experimental Psych.: Learning Memory Cognition*, Vol. 10, pp. 298-303.

H. Matsumoto, S. Hiki, T. Sone and T. Nimura (1973), "Multidimensional representation of personal quality of vowels and its acoustical correlates", *IEEE Trans. Audio Electroacoust.*, Vol. AU-21, pp. 428-436.

- T. Murry and S. Singh (1980), "Multidimensional analysis of male and female voices", *J. Acoust. Soc. Amer.*, Vol. 68, pp. 1294-1300.
- T. Murry, S. Singh and M. Sargent (1977), "Multidimensional classification of abnormal voice qualities", *J. Acoust. Soc. Amer.*, Vol. 61, pp. 1630-1635.
- G. Papcun, J. Kreiman and A. Davis (1989), "Long-term memory for unfamiliar voices", *J. Acoust. Soc. Amer.*, Vol. 85, pp. 913-925.
- I. Pollack, J.M. Pickett and W.H. Sumbly (1954), "On the identification of speakers by voice", *J. Acoust. Soc. Amer.*, Vol. 26, pp. 403-406.
- SAS Institute (1983), *SUGI Supplemental Library User's Guide* (SAS Institute, Inc., Cary, NC).
- H. Saslove and A. Yarmey (1980), "Long-term auditory memory: Speaker identification", *J. Appl. Psych.*, Vol. 65, pp. 111-116.
- S. Schiffman, M.L. Reynolds and F.W. Young (1981), *Introduction to Multidimensional Scaling: Theory, Method and Applications* (Academic Press, New York).
- S. Singh and T. Murry (1978), "Multidimensional classification of normal voice qualities", *J. Acoust. Soc. Amer.*, Vol. 64, pp. 81-87.
- J.A. Swets and R.M. Pickett (1982), *Evaluation of Diagnostic Systems: Methods From Signal Detection Theory* (Academic Press, New York).
- W.D. Voiers (1964), "Perceptual bases of speaker identity", *J. Acoust. Soc. Amer.*, Vol. 36, pp. 1065-1073.
- B.E. Walden, A.A. Montgomery, G.J. Gibeily, R.A. Prosek and D.M. Schwartz (1978), "Correlates of psychological dimensions in talker similarity", *J. Speech Hearing Res.*, Vol. 21, pp. 265-275.