Talker Recognition in Tandem with Talker-Independent Isolated Word Recognition

AARON E. ROSENBERG, FELLOW, IEEE, AND KATHLEEN L. SHIPLEY

Abstract—**A talker recognition system operating in tandem with a talker-independent isolated word recognizer is described and evaluated. The word recognizer uses a small set of reference templates for each vocabulary word. Each set is intended to span and typify individual talker templates over a large population of talkers. Word recognition decisions are based on template distance scores obtained by comparing processed input utterances to each set of reference templates. The distribution of distance scores for the templates corresponding to the actual word input has been found to be reasonably consistent for individual talkers, and to vary sufficiently from talker to talker to provide the basis for a talker recognition capability. A system has been implemented to exploit this capability. An evaluation of the system, carried out using a 100-talker database of digit utterances, shows that good talker recognition performance can be obtained for input utterances consisting of sequences of seven or more digits. Identification error rates varying from 3.6 to 14.0 percent for talker populations varying from 10 to 100 talkers are obtained. When the recognizer orders the talkers as candidates for recognition, the correct talker is found, on the average, among the top 0.8 percent of the population. Tested in a talker verification mode, the average error rate is approximately 8 percent.**

I. INTRODUCTION

In this paper we describe a talker recognition system which operates in tandem with a talker-independent isolated word recognition system. The word recognizer serves as the "front end" to the talker recognizer. The talker recognition capability is obtained with little additional computation than what is already required for the word recognizer and with only modest requirements for storage of talker prototype information.

There appears to be very little explicit discussion in the published literature of combined automatic speech and talker recognition. Two earlier studies that are related to the present work can be cited. Calavrytinos *et al.* [1] (abstract) investigated a combined speech and talker recognition process using a parametrization of the speech signal based on word-sized units. Improved talker recognition performance is reported for sequences of words over individual words. Kashyap [2] explored the possibilities of combined speech and talker recognition using phoneme-sized units. The use of sequences of phonemes to improve talker recognition performance and the grouping of talkers were also considered.

A simple approach to combining talker recognition with speech recognition is possible for talker-dependent, template-based isolated word recognizers. The prototype templates established for each talker for a given vocabulary in these systems are designed to characterize as uniquely as possible the words in the vocabulary. However, by the very nature of talker-dependent systems, for each word in the vocabulary the templates contain information characteristic of the talker. Therefore, it should be possible to operate such a system in a talker recognition mode by comparing analyzed input words from an unknown talker with templates for the input words across a population of talkers. In a talker verification mode, an identity claim is made, and each processed input word from the unknown talker is compared to the prototype word associated with the claimed identity. If the comparison results in a distance measurement that is less than a preassigned threshold, the identity claim is accepted.

Some commercial speech recognition systems offer a talker verification capability as an option. Although not explicitly stated, it seems reasonable to assume that such systems operate in the manner just described. Also, the performance of such systems in a talker verification mode has not been described. Although the templates used for word recognition are not explicitly designed for talker recognition, reasonable performance can be expected in talker recognition modes due to the talker-dependent nature of the templates. A study is now being undertaken to assess the performance of this kind of system.

Following is an overview of the proposed system for combining talker recognition with talker-independent word recognition. The input to this sytem is an unknown sequence of words from a specified vocabulary, spoken as isolated utterances by an unknown talker. The output provided by the system is identification of the words and a recognition decision (either identification or verification) concerning the talker. Identification of the words is carried out by an existing talker-independent word recognizer [3], while the talker recognition process makes use of distance or pattern similarity scores provided by the recognizer in the manner described below.

An overall block diagram of the system is shown in Fig. 1, with the word recognizer occupying the top portion of the figure and the talker recognizer the bottom. (The talker recognizer is shown in an identification mode.) Consider the input of a single word to the system. The "front-end" acoustic processor produces a parameterized version of the utterance, a pattern, which is compared to a set of word reference prototype patterns. For each word in the vocab-

Manuscript received June 14, 1984; revised November 14, 1984.

The authors are with the Acoustics Research Department, AT&T Bell Laboratories, Murray Hill, NJ 07974.

0096-3518/85/0600-0574$01.00 © 1985 IEEE

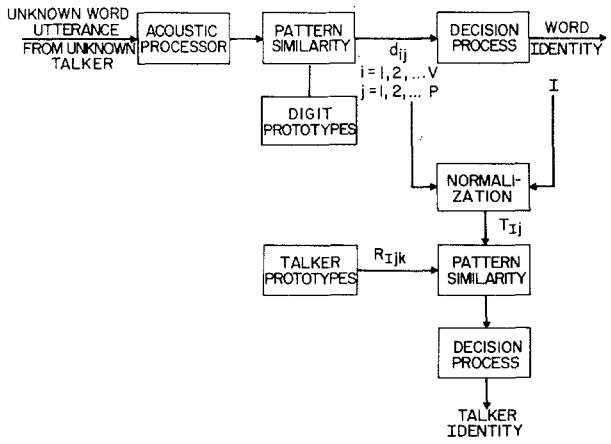

Fig. 1. Combined talker-independent digit recognizer and talker recognition system.

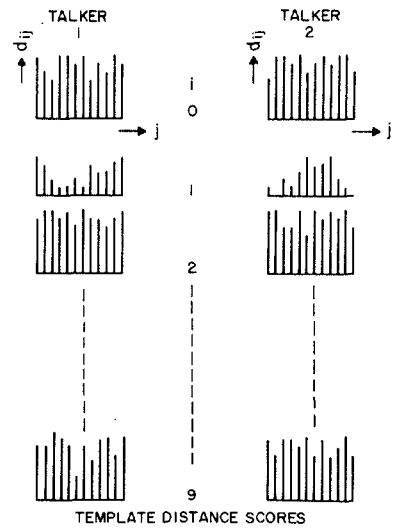

Fig. 2. Hypothetical sets of template distance scores for two talkers uttering the digit "1."

ulary there are P prototype patterns. Each set of P patterns is obtained by a statistical clustering procedure to typify and span individual talker patterns for the word over a large population of talkers [3]. The pattern similarity process outputs a distance score (or measure of dissimilarity) for each comparison of the input pattern with the set of reference patterns. The word identity decision is made as follows. Let $[d_{ij}]$ be the matrix of distance scores output by the pattern similarity process, where $i = 1, 2, \cdots, V, j = 1, 2, \cdots P$, and V and P are the number of words in the vocabulary and the number of prototype patterns per word, respectively. For each word i, the distance scores are sorted from lowest to highest (best to worst), such that

$$d_{ij_1} \leq d_{ij_2} \leq \cdots \leq d_{ij_P}. \tag{1}$$

Then the average of the K best (lowest) scores for each i is calculated as

$$r_i = \frac{1}{K} \sum_{q=1}^{K} d_{ij_q} \tag{2}$$

and the decision for the word recognized is

$$I = \underset{i}{\operatorname{argmin}}[r_i].$$

This represents the so-called K-nearest neighbor (KNN) decision rule [3]. For the talker-independent word recognizer used here, P is set to 12, and K is typically 1 or 2. In what follows we restrict attention to the vocabulary of spoken digits for which the word index i is renumbered $0, 1, \cdots, 9$.

We now present an example which shows how the digit identity decision is made and how talker information is contained in the reference prototype distance scores. A hypothetical set of distance scores for each of two talkers uttering the digit "1" is shown in Fig. 2. The height of each vertical line represents the magnitude of each (reference prototype) distance score, where j, the prototype index, runs along the abscissa, and the word index i is shown with each group of $P = 12$ distance scores, arranged vertically for each talker. By inspection, it can be seen that after sorting the distance scores and invoking the KNN rule with K equal to, say, 2, that the same correct decision, namely "1," would be obtained for each talker. However, the example is meant to illustrate the possibility that the distribution of scores as a function of j for the correct digit can be quite different for two talkers. For some of the prototypes for which small scores are obtained for talker 1, large scores are found for talker 2, and vice versa. The reason for this is that the clustering process establishes prototypes which are representative of distinct subpopulations of the training population of talkers. Thus, each of these test talkers can be expected to match some of these subpopulations well and others poorly, but not necessarily the same ones or to the same degree. Thus, the distance scores for the input word as a function of the prototype index j can be considered as a vector or profile of the talker's match to the clustered prototypes of the training population. Talker discrimination might be carried out by comparing the vectors or profiles of prototype distance scores of two talkers uttering the same word, thereby providing a talker recognition capability in tandem with talker-independent word recognition. Two conditions are necessary in order to obtain such a talker recognition capability, an intertalker condition and an intratalker condition. The intertalker condition is that the prototype patterns for each vocabulary word must sufficiently encompass and represent all expected "pronunciations" for the word. Thus, the training set of talkers must be large and comprehensive, the clustering procedure effective, and P sufficiently large. The current training set consists of 100 talkers, 50 male and 50 female, and P is fixed at 12. Under these conditions, word identification performance is comparable to identification performance in a talker-dependent mode with two prototypes per word. The intratalker condition is that for repeated utterances of the same word by a given talker, the distance scores profiles should vary little. In the following situation the condition does not hold. Suppose there are two (or more) com-

mon pronunciations of a word, and suppose they are represented in the prototype patterns for the word. If a talker uses more than one of these pronunciations in repeated utterances of the word, then the distance score profiles could change radically. A way of dealing with this possibility will be explored later in this paper.

In the remainder of the paper we will describe techniques for carrying out talker recognition tasks based on the hypothesis given above. In addition, an evaluation is presented to provide an indication of how well the system can perform.

II. Talker Recognition Process

As shown in Fig. 1, given the distance score matrix $[d_{ij}]$ and the identity of the input spoken digit, I, a normalization is carried out, resulting in a talker distance profile vector for digit I whose jth component is given by

$$T_{Ij} = \min [d_{Ij} - d_{Ij\min}, T_{\max}] \tag{3}$$

where

$$j_{\min} = \underset{j}{\operatorname{argmin}} [d_{Ij}]. \tag{4}$$

Subtraction of the minimum distance score compensates for a constant bias added to all the scores for a given utterance. The constant $T_{\max}$ is used to restrict the range of profile values.

Talker prototypes consist of the mean and variance of test talker profiles obtained from a set of training utterances (typically five) for each vocabulary word and each talker. Thus, the mean is

$$R_{Ijk} = \frac{1}{N} \sum_{n=1}^{N} T_{Ijkn} \tag{5}$$

and the variance is

$$s_{Ijk}^2 = \max \left\{ \frac{1}{N} \sum_{n=1}^{N} (T_{Ijkn} - R_{Ijk})^2, s_{\min}^2 \right\} \tag{6}$$

where T_{Ijkn} is the jth component of the test profile for the nth training utterance for digit I from talker k, and N is the number of training utterances. $s_{\min}$ is a constant which is used to control the relative influence of the components in a weight vector derived from the s's, used in the talker distance computation described further on. The jth component of the weight vector is given by

$$w_{Ijk} = \frac{s_{Ijk}^{-1}}{\sum_{j'=1}^{P} s_{Ij'k}^{-1}}. \tag{7}$$

As $s_{\min}$ becomes very large, the weight components all tend to $1/P$, that is, there is no relative weighting at all.

The metric used in the pattern similarity process shown in Fig. 1 is a weighted "city block" metric. It provides a talker distance in the comparison of a test profile from an unknown talker for digit I with the reference profile from talker k for the same digit.

$$D_{Ik} = \sum_{j=1}^{P} w_{Ijk} |T_{Ij} - R_{Ijk}| \tag{8}$$

where w_{Ijk} is the jth component of the weight vector defined in (7).

In general, the unknown talker provides a sequence of M digit utterances $S, I_1, I_2, I_3, \cdots, I_M$, and an accumulated talker distance D_{Sk} is calculated,

$$D_{Sk} = \sum_{m=1}^{M} D_{I_m k}. \tag{9}$$

The talker distance is used to provide a decision in various talker recognition tasks. In talker identification, test profiles of the unknown talker are compared to reference profiles for each talker in a population of K talkers. The identified talker is taken to be

$$k_{\min} = \underset{k}{\operatorname{argmin}} [D_{Sk}]. \tag{10}$$

In talker verification, the unknown talker provides an identity claim k, in addition to a sequence of digit test utterances. The talker's test profiles are compared against reference profiles for talker $\hat{k}$, resulting in a talker distance $D_{S\hat{k}}$. This is compared to a talker distance threshold $\tau_{S\hat{k}}$, which is obtained from the reference information. If $D_{S\hat{k}} < \tau_{S\hat{k}}$, then the identity claim is accepted; otherwise it is rejected.

In both talker identification and talker verification, it is possible to set additional thresholds associated with a "no decision" or deferred decision category. Such thresholds are used to lessen the risk of making an incorrect decision. In talker identification, this decision category is required in the "open-set" situation, in which it is not known whether the unknown talker is included in the reference population.

III. Two-Class References

It was noted earlier that intratalker variations can produce significantly different talker profiles for utterances of the same word. This situation would offer no difficulty for talker recognition if there were a reference profile for each pronunciation of a given word for each talker. Of course, there is no way to guarantee that all of a talker's pronunciation modes will occur in a finite set of training utterances. Moreover, as a practical consideration, the number of both training utterances and reference profiles should be kept as small as possible.

In this study the following method was used in consideration of the possibility of needing more than one reference profile per word per talker. The number of training utterances per word remains fixed, but the profiles for these utterances are compared to each other using a hierarchical clustering procedure [4] to identify classes of talker profiles which differ significantly. The procedure starts by considering each training utterance profile as a class and the distance between two classes as the talker distance between the two profiles. Then, classes are merged iteratively, using as the distance between two classes the min-

imum of all the talker distances between the two classes, until two classes remain. The distance between these two classes, called D_{sep}, is used to determine whether the classes differ sufficiently to form two reference profiles from the training utterances instead of one.

For two-class reference profiles, the test talker distance D_{Ik} is assigned to the minimum of the two class talker distances, as follows:

$$D_{Ik} = \min_{c}[D^c_{Ik}] \tag{11}$$

where

$$D^c_{Ik} = \sum_{j=1}^{P} w^c_{Ijk} \, |T_{Ij} - R^c_{Ijk}| \tag{12}$$

and R^c_{Ijk} and w^c_{Ijk} are the jth components of the cth class of reference profile vectors and weights for talker k and word I, and $c = 1$ or 2.

IV. Thresholds and Errors in Talker Verification

In the talker verification mode, the talker distance is compared to a threshold to determine whether to accept or reject an identity claim. Two kinds of error are possible. If the talker's identity claim is correct but is rejected by the system, a false rejection is said to occur. If the talker's identity claim is false but is accepted by the system, a false acceptance occurs. Suppose the talker distance is considered to be a random variable. Then the probability of false rejection and the probability of false acceptance are given by

$$P_{\text{FR}} = \text{Prob}\{D_{Sk} > \tau_{Sk}\} \tag{13}$$

and

$$P_{\text{FA}} = \underset{l \neq k}{\text{Prob}}\{D_{Sl} < \tau_{Sk}\} \tag{14}$$

where k represents the talker whose identify is claimed and l represents any other talker. τ_{Sk} is the threshold distance appropriate for talker k and digit sequence S. For here, and for what follows, to simplify the notation, the subscript k refers to talker self-distance, whereas subscript l refers to distances for any talker other than k with k's reference profiles.

Thresholds can be specified by estimating the probability distribution of D_{Sk} or the joint probability of all D_{Sl}'s and choosing a threshold value associated with a desired probability of false rejection or false acceptance. In practice, it is usually simpler to set a threshold based on probability of false rejection using talker distances obtained from a talker's own training utterances. Unfortunately, such estimates are often quite poor, due to the usually small number of training utterances available. Also, since the training profiles are used to create reference profiles, these talker distances are biased estimates of the true talker distance. Consequently, it is advisable to use training profiles only to provide initial threshold estimates, and to use actual test profiles to correct and adapt these estimates. Adaption will be described in a later section.

Threshold computation in the system described in this report is complicated still further by the fact that an input utterance S, in general, consists of a sequence of digit utterances. Talker distances are generally different for each word, and are not necessarily statistically independent of each other. Thus, errors in threshold estimates can accumulate rapidly as the length of the sequence of utterances increases. It will be shown that under these circumstances threshold adaption is important.

Let

$$\mu_{Sk} = \text{E}\{D_{Sk}\} = E\left\{\sum_{m=1}^{M} D_{I_m k}\right\} = \sum_{m=1}^{M} \mu_{I_m k}$$

and

$$\begin{aligned} \sigma^2_{Sk} &= \text{var}\{D_{Sk}\} = \text{var}\left\{\sum_{m=1}^{M} D_{I_m k}\right\} \\ &= \sum_{m=1}^{M} \sigma^2_{I_m k} + 2 \sum_{m<n} \rho_{I_m I_n k} \, \sigma_{I_m k} \, \sigma_{I_n k} \end{aligned} \tag{16}$$

be the means and variances of the talker distance of the sequence S for talker k in terms of the means, variances, and combinations of the talker distances for the digit utterances making up the sequence. Then, given normally distributed talker distances, the probability of false rejection is given by

$$\text{Prob}\{D_{Sk} > \tau_{Sk}\} = 1 - \Phi\left[\frac{\tau_{Sk} - \mu_{Sk}}{\sigma_{Sk}}\right] \tag{17}$$

where $\Phi(\cdot)$ is the cumulative normal distribution. This probability is the area under the upper tail of the distribution. If we require that

$$\text{P}_{\text{FR}} = \alpha \tag{18}$$

then we set

$$\tau_{Sk} = \mu_{Sk} + C_\alpha \sigma_{Sk} \tag{19}$$

where C_α is a constant such that

$$\alpha = 1 - \Phi(C_\alpha). \tag{20}$$

In this study, initial estimates are obtained for μ_{Ik} for each digit I and talker k, and allowed to adapt through a sequence of test utterances, as described in Section V. However, compromise estimates are used for σ_{Ik}, in one case, independent of I, and in the other case, independent of both I and k. In addition, we assume that $\rho_{I_m I_K}$ is zero. This should be a reasonably good assumption except for the case of repeated digits. With these assumptions, (19) becomes

$$\tau_{Sk} = \sum_{m=1}^{M} \mu_{I_m k} + \text{C}_\alpha \sqrt{M} \sigma_k \tag{21}$$

for the case of $\sigma_{Ik} = \sigma_k$, and

$$\tau_{Sk} = \sum_{m=1}^{M} \mu_{I_m k} + C_\alpha \sqrt{\text{M}} \sigma \tag{22}$$

when $\sigma_{Ik} = \sigma$ for all I and k.

V. Adaptation

In talker recognition applications where talkers access the system repeatedly over periods of time, it is advisable to consider adaptation or updating reference data. Typically, talker verification tasks involve sequences of trials over long periods of time for each talker, whereas for talker identification, infrequent or unique comparisons take place for a single talker. In talker verification, updating both reference profile and threshold parameters with information from the test utterance of successful (accepted) trials is useful to compensate and improve initial estimates, which, as has been mentioned, are often poor due to insufficient or faulty training data, as well as to track long-term variations which may occur for many talkers. In this study, we will evaluate the use of adaptation of both reference profile and threshold data in simulations of talker verification trials.

It is supposed that a trial occurs in which a sequence of M digits, S, is uttered by talker k. For adaption to take place, it is necessary that

$$D_{Sk} < \tau_{Sk} \tag{23}$$

where D_{Sk} is defined in (9) and τ_{Sk} is given in (21) or (22).

The general procedure for adapting a parameter p is as follows. Let p_{old} and p_{new} be the previous and adapted values of the parameter, and $\hat{p}$ an estimate of the parameter obtained from the current trial. Then

$$p_{\text{new}} = \gamma\hat{p} + (1 - \gamma)p_{\text{old}} \tag{24}$$

where

$$\gamma = \max\left(\frac{1}{T}, \gamma_{\min}\right), \tag{25}$$

T is a count of the number of adaptations of the parameter (including the current one), and $\gamma_{\min}$ is typically 0.25. Thus, the adaptation is a weighted average of the current estimate and previous value of the parameter.

For threshold parameter adaptation, $\mu_{I_{mk}}$ in (21) and (22) is estimated by the current value of $D_{I_{mk}}$. σ_k in (21) is estimated by the sample variance of the $D_{I_{mk}}$. Initial estimates are obtained from the training data.

Two adaptation procedures have been tried. In procedure 1, the formulation for τ_{Sk} given in (21) is used along with the criterion expressed by (23). The threshold μ's and σ's as well as the reference profiles R's and s^2's are adapted.

In procedure 2, the formulation given in (22) is used. In addition, the adaptation criterion (23) is expanded as follows. If

$$D_{I_{mk}} < \mu_{I_{mk}} - B \tag{26}$$

then the μ's, R's, and s^2's are adapted as described above. If (23) is satisfied but not (26), then only the μ's are adapted as

$$(\mu_{I_{mk}})_{\text{new}} = D_{I_{mk}} + B. \tag{27}$$

The constant B is a "security" bias applied to the estimated means. The effect of this kind of updating with a bias is to tend to maintain μ_{Ik} at a value B greater than immediate past values of D_{Ik}. Since biased estimates of the means are obtained, the threshold effectively includes a term that grows linearly with sequence length, in addition to the term that grows with the square root.

VI. Experimental Evaluation

In order to carry out an evaluation of the capability of the talker recognition technique described in the first part of this report, a database of spoken digits was collected from a population of 100 talkers, 50 male and 50 female. Each talker provided 20 utterances of each digit in five recording sessions over a period of approximately 2 weeks. Four sets of digits were recorded in each session with the digits selected randomly in each set. Utterances were recorded over dialed-up telephone lines through a local exchange with the talker seated in a sound booth. An ordinary carbon button telephone handset was used. All utterances in a session were recorded in a single call. Each utterance was processed two ways. For the primary processing, utterances were passed through the real-time front end of the word recognizer whose output is a vector of autocorrelation coefficients (eighth order) every 15 ms. The recognizer endpoint detector was used to mark the beginning and end of each utterance. However, endpoints were monitored and manually adjusted, if necessary, in the event of an apparent error. 320 such corrections were required out of the 20 000 utterances.

The database was established as endpointed arrays of autocorrelation coefficient vectors. In this form, the utterances were then passed through the comparison block of the talker-independent digit recognizer to provide a distance score matrix for each utterance. These, in turn, together with the identity of the input digit, were processed according to (3) to provide distance score profiles. Note that the distance score profiles correspond to the input digit. This is not necessarily the digit identified by the recognizer. In other words, the digit identity provided by the recognizer is ignored, so that digit identification in the evaluation is assumed to be perfect.

In the experiments reported here, the profiles corresponding to the first five utterances of each digit were designated as training profiles and were used to construct reference profiles for each talker according to (5) and (6). The remaining profiles were designated test profiles. All profiles were selected in the order in which the utterances were recorded.

Since talker population size is a variable of interest in the experiments, the 100 talkers were assigned to talker subsets of various sizes. There are 20 subsets of size 10, 10 subsets of size 20, and four subsets of size 50, in addition to the population as a whole. Talkers were selected at random for each subset, with the possibility of talkers assigned to more than one subset of a given size.

The basic trial for the experiments consists of the sim-

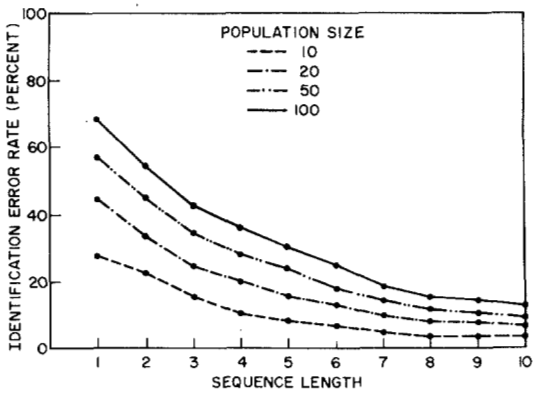

Fig. 3. Average identification error as a function of sequence length for the four talker population sizes (type 1 digit sequence, $s_{min} = 0.05$, two-class references).

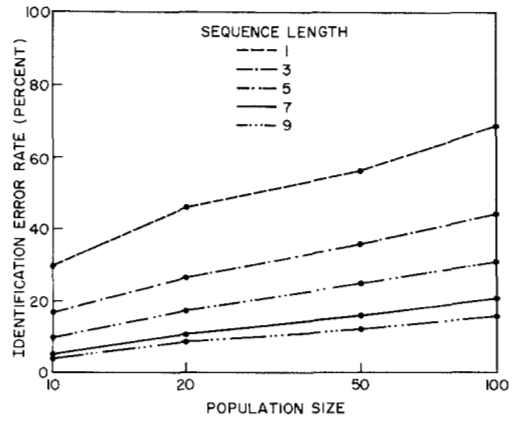

Fig. 4. Average identification error as a function of talker population size for five sequence lengths (type 1 digit sequence, $s_{min} = 0.05$, two-class references).

ulation of a sequence of one to ten digit utterances. Test profiles were selected corresponding to the digits which are prescribed for the sequence and compared to reference profiles for each digit. The resulting distances were accumulated for use in either a talker identification or talker verification mode. Errors were tabulated over the talkers in a subset to provide performance data. Both the length and composition of the digit sequences are experimental variables.

A. Talker Identification Experiments

For evaluation of talker identification performance, an experimental series was carried out in which the basic experiment consisted of a series of ten trials for each talker, the first trial consisting of a single digit utterance, with each succeeding trial adding an additional digit utterance up to the tenth trial, which has ten digit utterances. For each selection of experimental variables, the experiment is repeated ten times, using a different set of test profiles for each experiment.

The results are in the form of average talker identification error rate as a function of the number of utterances in a trial (sequence length) and size of the talker subset. The experimental parameters which are varied from one experimental series to another are the composition of the digit sequences, a parameter τ_{sep}, controlling the number of reference classes per talker, and the value of s_{min} which controls the weights in the talker distance calculation [see (6)–(8)].

Four kinds of digit sequence compositions have been examined. These are type 1, a fixed sequence 0, 1, 2, $\cdots$, 9; type 2, random sequences of digits with no duplications allowed; type 3, random sequences with duplications allowed; and type 4, sequences of repeated digits.

Typical results are shown in Fig. 3. In this case the fixed digit sequence is used, s_{min} is set at 0.05,[1] and two-class references are used. As a function of sequence length, there is an initial rapid improvement in error rate followed by a leveling off. There is a suggestion that the leveling off begins at shorter sequence lengths for smaller population sizes. For single utterance sequences, error rates range from approximately 30 percent for ten-talker populations to 70 percent for the entire 100-talker population. For sequences of length 8 or more, the range is from 3.5 to 12.5 percent.

The improvement in performance that occurs as sequence length increases can be expected if the distributions of talker distance from utterance to utterance in a sequence are relatively statistically independent. As shown in (15) and (16), as sequence length M increases, mean distance increases linearly with M, while the standard deviation increases with the square root if the distances between utterances are uncorrelated. Thus, talker distance distributions over such sequences become sharper with increasing M, providing improved discrimination between talkers. In cases in which statistical independence cannot be expected to hold well, as in sequences with repeated utterances, performance improvement is also not expected, as shown further on. The same results are plotted as a function of population size for five different sequence lengths in Fig. 4. If the population size is scaled logarithmically, then the error rate increases almost linearly.

Variability in performance from talker to talker is an important aspect of these results. Histograms of individual talker error rates are shown in Fig. 5 for sequence lengths 1, 3, 5, 7, and 9. It can be seen that the histograms are quite broad, and, in particular, as sequence lengths increase and average error rates decrease, the histograms become considerably skewed to the right with significant tails.

A more complete analysis of identification error results takes into account the rank of the correct talker. This is the number of talkers which score better in an identification trial than the correct talker. Average rank results corresponding to the identification error results shown in Fig. 3 are displayed in Fig. 6. The data are plotted as average

[1]For reference, the average value and median of the standard deviation of profile distances over training utterances [the square root of the first term in brackets in (6)] is approximately 0.06.

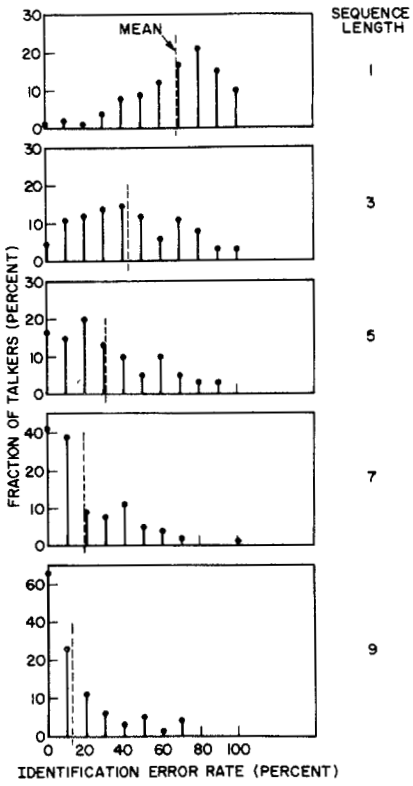

Fig. 5. Histograms of the distribution of individual talker error rates for five sequence lengths (type 1 digit sequence $s_{min} = 0.05$, two-class references).

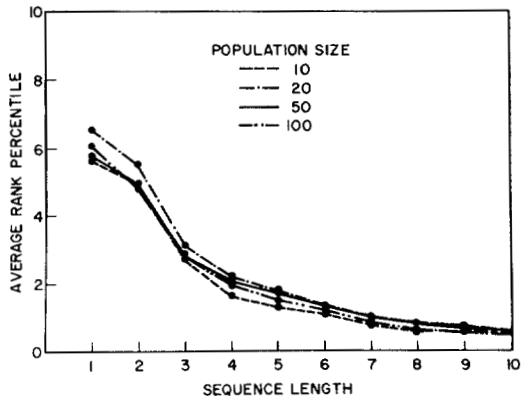

Fig. 6. Average identification rank percentile as a function of sequence length of four talker population sizes (type 1 digit sequence, $s_{min} = 0.05$, two-class references).

rank percentile as a function of sequence length for the four talker subset sizes. Rank percentile is the fraction of the talker population (in percent) which ranks better than the correct candidate. It can be seen that average rank percentile is essentially independent of population size. It drops off rapidly as a function of sequence length to approximately 0.6 percent for sequence lengths 8 or greater. Thus, although absolute identification error can attain high values, the correct candidate ranks quite well on the average.

The effect of different sequence compositions on error rate is shown in Fig. 7. These results are for the case in which population size is 20, s_{min} is 0.05, and references

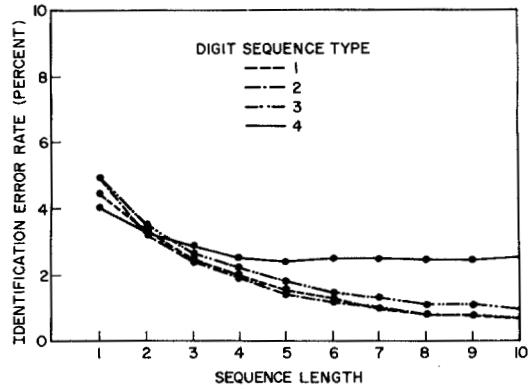

Fig. 7. Average identification error rate as a function of sequence length for four digit sequence composition types for talker population size of 20 ($s_{min} = 0.05$, two-class references).

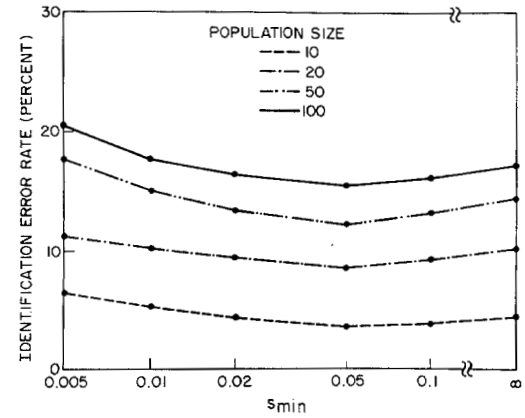

Fig. 8. Average identification error rate as function of s_{min} for four talker population sizes; results are averaged over sequence lengths 8, 9, and 10 (type 1 digit sequence, one-class references).

are two-class. The best overall performance is obtained for fixed digit sequences and random sequences without duplication. Both of these sequences have no duplicated digits. Since the difference in composition between these sequences vanishes as sequence length approaches 10, as can be seen, so does the difference in error rate. Error rates are uniformly greater for random sequences with duplication allowed. For such sequences, the correlation between utterances of the same digit produces larger variances in the talker distance distributions. The extreme situation is for sequences composed solely of repeated digits. In this case there is no improvement in error rate for sequence lengths greater than 5, and the error rate does not fall below 20 percent.

To examine the usefulness of applying weights, derived from the variance over the set of training profiles, in the talker distance formulation, (8), the parameter s_{min} is varied. The results are shown plotted in Fig. 8. These results are obtained for fixed digit sequences and one-class references. The error rates shown are obtained as a result of averaging error rates for sequence lengths 8, 9, and 10, and are shown for each of the four talker subset sizes. Six values of s_{min} were evaluated, five values ranging from 0.005 to 0.1, plus a very large value equivalent to no weighting. The error rates are shown as a function of s_{min}

TABLE I
AVERAGE IDENTIFICATION ERROR RATES (PERCENT) AS A FUNCTION OF TALKER POPULATION SIZE FOR THE BEST EXPERIMENTAL VALUE OF THE DISTANCE WEIGHTING PARAMETER s_{min} AND FOR NO WEIGHTING ($s_{min} = \infty$)

size	s_{min}	
	0.05	∞
10	3.63	4.52
20	8.62	10.25
50	12.25	14.47
100	15.60	17.30

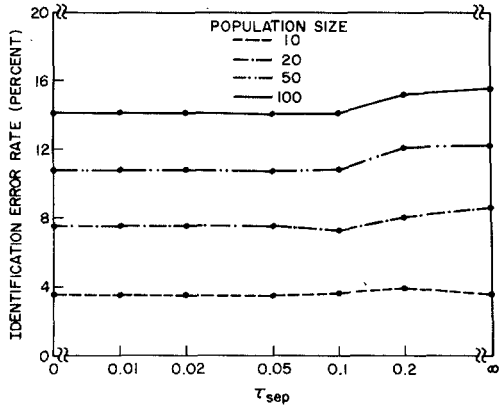

Fig. 9. Average identification error rate as a function of τ_{sep}, threshold distance between reference classes, for four talker population sizes; error rates are averaged over sequence lengths 8, 9, and 10 (type 1 digit sequence, $s_{min} = 0.05$).

scaled logarithmically with rates for the unweighted condition plotted on the right-hand axis. It can be seen that there is an optimal value of s_{min} equal to 0.05, for which there is a small but consistent improvement in average error rate of the order of 1 or 2 percent over the unweighted condition. The actual error rates are shown in Table I.

Finally, experiments were carried out to compare the effectiveness of using two-class references and one-class references. Actually, a more general test was carried out varying a parameter τ_{sep}, which controls (on a talker and digit basis) whether one- or two-class references are used. D_{sep}, as described in Section III, is the distance betwen the two classes of profiles. Suppose a threshold τ_{sep} is applied such that if D_{sep} for a given talker and digit falls below the threshold, the one-class reference profile is used; otherwise, the two-class reference profile is used. It is hypothesized that there is some intermediate value of τ_{sep} associated with a mixture of one-class and two-class reference profiles which leads to improved performance over either all one-class or two-class references. The reasoning behind this hypothesis is the following. The hierarchical clustering procedure provides two-class reference profiles for every talker and digit. However, two reference profiles may not be appropriate for every talker and digit. In some cases the profiles making up the training set may be quite uniform, so that fragmenting them into two reference profiles is unnecessary and perhaps harmful. The harmful aspects are the loss in statistical efficiency in estimating the profiles and the effective increase in the population of reference profiles with an increased potential for misidentification. The results are plotted in Fig. 9 for the case of

TABLE II
AVERAGE IDENTIFICATION ERROR RATES (PERCENT) AS A FUNCTION OF TALKER POPULATION SIZE FOR ONE-CLASS AND TWO-CLASS REFERENCES AND THE BEST EXPERIMENTAL VALUE OF τ_{sep} FOR EACH SIZE

size	1-class	2-class	best τ_{sep}
10	3.63	3.57	3.52
20	8.62	7.57	7.30
50	12.25	10.77	10.70
100	15.60	14.13	14.03

fixed digit sequences and s_{min} set to 0.05. Again, the error rates are obtained by averaging the error rate for sequence lengths 8, 9, and 10, and are shown for each of the four talker subset sizes. Error rates are plotted as a function of τ_{sep} scaled logarithmically, except that one-class results ($\tau_{sep} = 0$) are shown on the left-hand axis and two-class references ($\tau_{sep} = \infty$) are shown on the right-hand axis. The results show that the difference in error rates between one- and two-class reference profiles is of the order of 1 or 2 percent. However, there is essentially no benefit to using an intermediate value of τ_{sep}. Actual error rates for one-class and two-class references, and the best value of τ_{sep} for each size (either 0.05 or 0.1), are shown in Table II.

B. Talker Verification Experiments

For talker verification, an experiment is defined as a series of trials, each of which has the same sequence length. Experiments are carried out for sequence lengths varying from one to ten utterances. The number of trials per experiment varies from a maximum of 30 for sequence lengths of five or fewer utterances to 15 for ten-utterance sequence lengths. Structured in this way, the trials can be considered a sequence over time of sample utterances to which adaptation can be applied and the effects on performance measured. The sample profiles for each experiments are drawn from the 150 available test profiles.

Three different kinds of digit sequence compositions are examined, types 2, 3, and 4, as described in Section V-A. For type 4, the number of trials is fixed at ten, each trial being a repeated sequence of a different digit. For this composition type, since each digit occurs in only one trial per experiment, there is no possibility of applying and observing the effects of adaptation. That is, it is not possible for the thresholds or reference profiles for any trial to have been modified in previous trials.

In each trial for each talker, the comparison of that talker's test profiles with his reference profiles is used to measure false rejection performance. The comparison of the remaining talkers' test profiles to the reference profiles is used to measure false acceptance performance. It is thus implicit here, as in most evaluations of this kind, that "impostors" are talkers who repeat the same utterances as the "true" talkers, but make no attempt to imitate the "true" talkers' voices.

Fig. 10 shows average and standard deviation false reject and false accept rates as a function of talker subset size, where the subsets are the same subsets used in the identification experiments. The results are shown for type

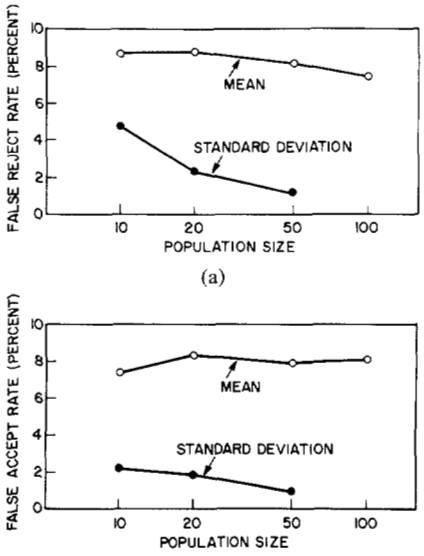

Fig. 10. Mean and standard deviation (a) false reject rate and (b) false accept rate as function of population size (type 2 digit sequence, sequence length 7).

2 digit sequence compositions of length 7. It can be seen that both false reject and false accept rates remain fairly constant. This result is not unexpected. The impostors are drawn from the same overall population so that the average accept impostor rate should be the same for any subset of that population. The standard deviation in each case decreases as population size increases. This trend is consistent with what would be expected from random samples as sample size grows. Less overall variation is seen in the false accept rate, since the number of "impostor" comparisons is approximately the square of the number of "true" talker comparisons.

Taking the foregoing results into consideration, in the remaining experiments verification performance is shown only for the total talker population size of 100.

First, the effects of digit sequence composition and length are considered. Performance is specified by the average *a posteriori* equal-error rate. For each talker, empirical cumulative talker distance distributions are obtained. Self-distances are obtained from comparisons of "true" test profiles to the reference profiles, while "impostor" distances are obtained from comparison of other talkers' test profiles to the reference profiles. Equal-error rate is associated with the distance for which the percentage of "true" distances greater than the distance is equal to the percentage of "impostor" distances less than the distance. The equal-error rates and distances are tabulated for each talker and averaged.

Average equal-error rates as a function of sequence length are shown in Fig. 11 for the three different types of digit sequence compositions. Type 1 compositions are omitted, since we have already seen from the talker identification results that type 1 compositions and type 2 compositions are essentially indistinguishable. The same trends that were apparent for talker identification in Fig. 7 are apparent here. For example, for type 2 compositions, single digit utterance error rate is approximately 18 percent. Performance improves with increasing sequence lengths to about 7 or 8 percent for lengths 7 or greater. For type 3 compositions, where duplication of digits is allowed, error rates are about 1 percent greater, and for type 4 compositions, all repeated digits, error rate is never better than 13 or 14 percent.

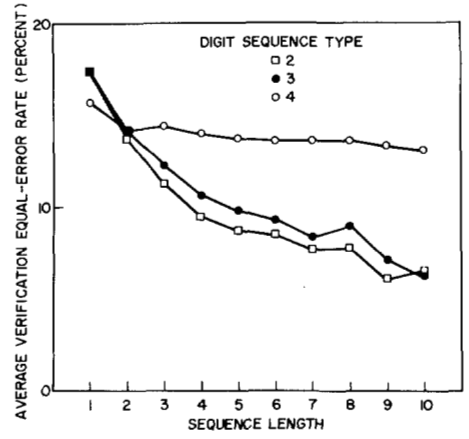

Fig. 11. Average verification equal-error rate as a function of sequence length for three sequence composition types (s_{min} = 0.05, two-class references).

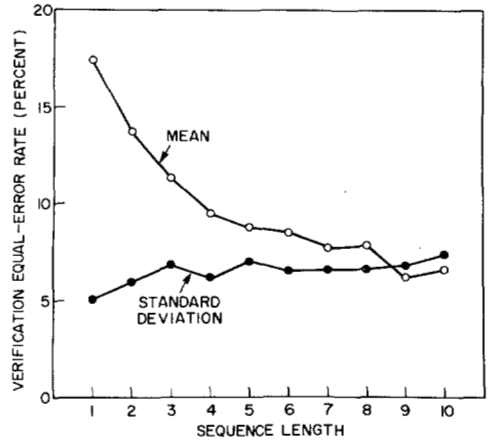

Fig. 12. Mean and standard deviation verification equal-error rate as a function of sequence length (type 2 digit sequence, s_{min} = 0.05, two-class references).

Variability in performance from talker to talker is a significant experimental outcome for talker verification, as it was for talker identification. The average equal-error rates for type 2 compositions shown in Fig. 11 are plotted again in Fig. 12, along with the standard deviation over the talker population. Even though average error rate decreases as a function of sequence length, standard deviation actually increases somewhat. This indicates the presence of a group of talkers with persistently bad performance, for whom little or no improvement in error rate is obtained with increasing sequence length. This aspect of talker variability can be seen more directly in the histograms of individual talker performance plotted in Fig. 13 for the same experiments. The fraction of the talker population falling in the

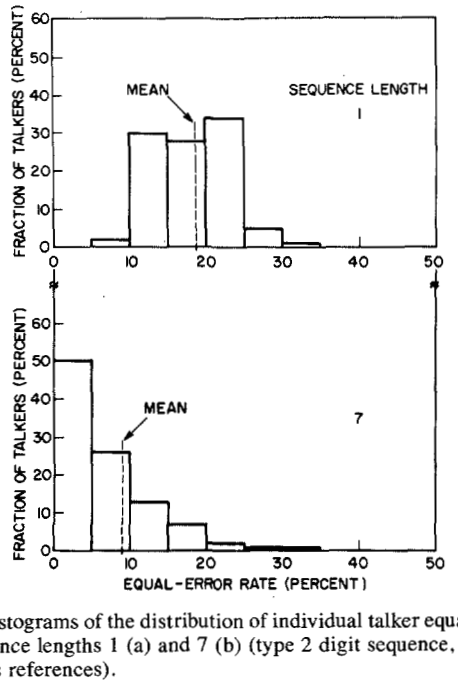

Fig. 13. Histograms of the distribution of individual talker equal-error rates for sequence lengths 1 (a) and 7 (b) (type 2 digit sequence, $s_{min} = 0.05$, two-class references).

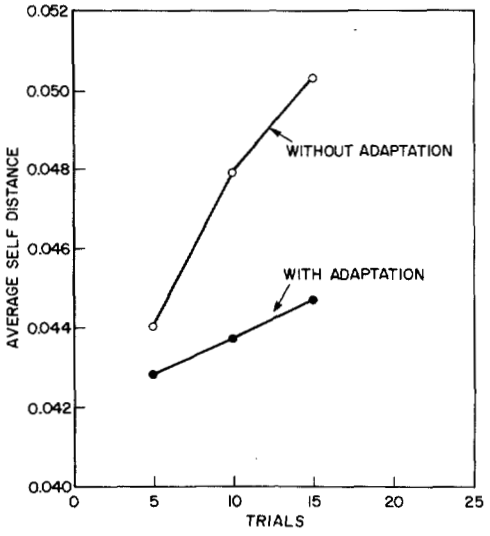

Fig. 14. Average talker self-distance as a function of number of (successive) trials with and without reference profile adaptation.

error rate ranges specified on the abscissa is shown. The top half of the figure shows the histogram for single utterance sequences, while the bottom half shows histograms for seven-digit-long utterances. Average error rate decreases significantly from one- to seven-digit-long utterances, but a small but persistent "tail" of bad performers with error rates greater than about 25 percent is apparent.

The foregoing results were obtained with two-class reference profiles and the distance weighting parameter s_{min} set at 0.05. For talker verification we will not consider the effects of varying these parameters, as we did for talker identification, since the results are quite similar. Instead, we turn now to tests of the threshold setting and threshold and reference profile adaptation procedures described in Sections V-A and V-B.

The talker verification results so far shown have been obtained by analyzing the cumulative distribution functions of "true" and "impostor" talker distances. In actual practice, performance must be considered relative to the techniques that are used for setting and adapting rejection thresholds. Errors are measured by tabulating the number of false rejections and false acceptances that occur with specified thresholds. As described in Section IV, a common approach for setting thresholds is to decide in advance on an allowable rate of false rejection and then estimate the appropriate threshold from available "true" distance sample distribution parameters. "True" distance samples may be obtained from training utterances and from test utterances in which "accept" decisions have been made. If the latter kinds of utterances are used, then adaptation is said to take place. Utilizing test utterance data to estimate thresholds is recommended not only for improving initial estimates from training data with additional samples, but also because of the often observed phenomenon that "true" talker distributions vary over time from trial to trial. This kind of behavior is illustrated in Fig. 14, where we have plotted average "true" or self-distance as a function of time over the period in which test utterances were recorded. Each data point represents the average self-distance over all talkers and recorded. Each data point represents the average self-distance over all talkers and (single) digit utterances for each successive group of five test utterances. (Reference profiles were constructed from five training utterances recorded preceding all the test utterances for each digit.) Results are shown without and with reference profile adaptation. Without reference profile adaptation, average self-distance increases at a much greater rate than with it. Since threshold estimates are proportional to the average self-distance [see (19)], they must adapt to larger values to maintain a constant rejection rate. Therefore, it is important to adapt reference profiles as well as thresholds to maintain good performance.

These same points are illustrated again by the error rate results shown in Fig. 15. Average error rates are shown as a function of sequence length (number of utterances) per trial for experiments in which thresholds were set and adapted using procedure 1. Two cases are shown—one with and the other without reference profile adaptation. The data points are obtained by tabulating false rejections and false acceptances, calculating average false rejection and false acceptance rates over all trials and talkers, and then averaging the rates. Also, the average equal-error rates shown in Figs. 11 and 12 are replotted again here. We find that for utterance lengths 4 or greater, there is a consistent 2 percent advantage in error rate associated with reference profile adaptation. (For smaller lengths, the adaptation is presumably not "fast" enough to provide a noticeable difference.) Equal-error rate results, for which there is no adaptation, are shown to provide a baseline performance comparison. The equal-error rate results are close to the results with both threshold and profile adaptation, and for some points, actually slightly worse.

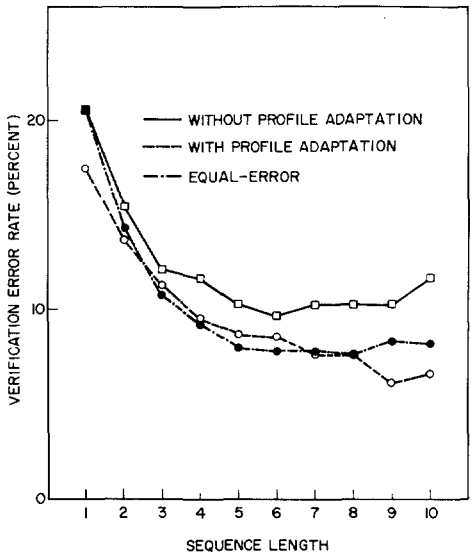

Fig. 15. Average verification error as a function of sequence length without, and with reference profile adaptation, and equal-error rate results (type 2 digit sequence, $s_{\min} = 0.05$, two-class references).

We now compare results for the two procedures described in Section V for setting and adapting thresholds. In particular, we examine their relative success at maintaining a constant customer reject rate independent of sequence composition or length. As indicated earlier, threshold setting and adaptation is greatly complicated for this system, because of the need for the threshold determination not only to be sensitive to individual talkers, but also to the composition and length of the sequence of digit utterances in each trial. (If talkers were expected to repeat the same utterances from trial to trial, much simpler empirical tracking procedures could be used.) Procedure 1 includes a straightforward formulation of threshold based on normal self-distance distributions, assuming uncorrelated utterances. Both means and standard deviations are updated using weighted averages of previous values and current distances. Procedure 2 assumes the same threshold formulation, but updates only the means and incorporates a bias in the estimate. Rejection rates are set by fixing C_α in procedure 1 and a combination of C_α and B in procedure 2. (In practice, for procedure 2, C_α was set at 1 and B varied to obtain a desired rejection rate.)

Results are shown for both procedures in Fig. 16. In both cases the parameters were set experimentally to obtain a normal 8 percent rejection rate. Fig. 16(a) shows results for procedure 1 and Fig. 16(b) for procedure 2. Average false reject and false accept rates are plotted as function of sequence length (number of utterances per trial) along with the average of the two. The average error rate plots are quite similar for each procedure. However, procedure 2 is much more successful at maintaining a constant reject rate independent of the number of utterances per trial. For procedure 2 the reject rate increases to approximately 13 percent (offset by a concomitant drop in false accept rate), for long sequences. We speculate that procedure 2 provides better results, because the updating procedure incorporating the bias reacts more sensitively to individual variations and trends in self-distance.

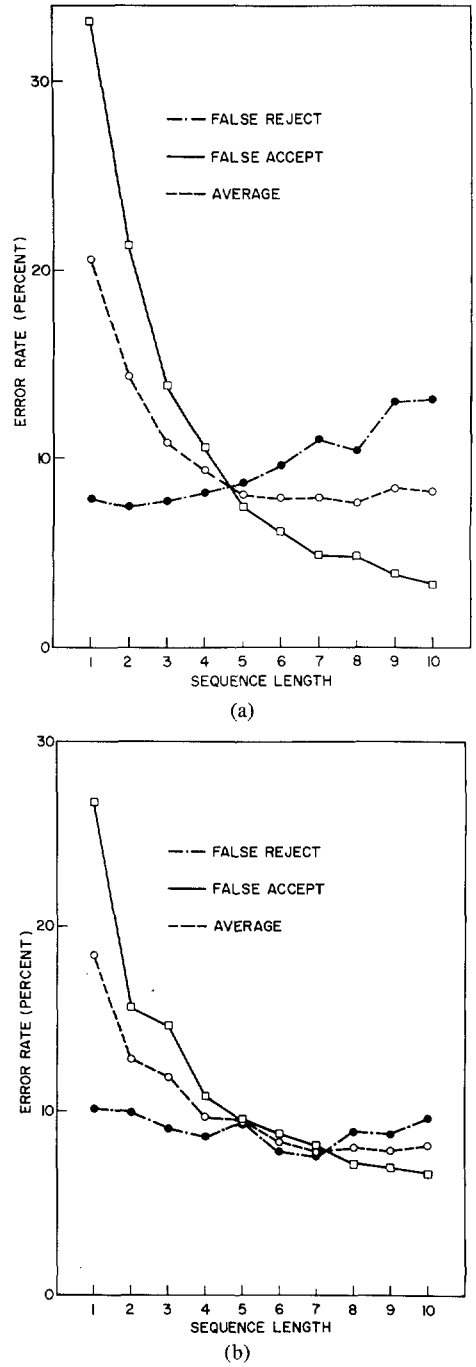

Fig. 16. Average false reject, false accept, and overall error rate as a function of sequence length for (a) adaptation procedure and (b) adaptation procedure 2 (type 2 digit sequence, $s_{\min} = 0.05$, two-class references).

VII. Discussion and Conclusion

In this report we have described a sytem that carries out automatic talker recognition tasks in tandem with speaker-independent isolated word recognition. This work is a followup on earlier work reported on in 1981 [5]. Some of the original techniques have been modified and a new, much larger database has been used to evaluate perfor-

mance. Modifications include elimination of a profile quantizing process, the addition of weights in the talker distance formulation, and the use of two-class reference prototypes. Special attention has been given to talker verification techniques, including problems involved in setting and adapting thresholds and adapting reference profiles for sequences of digit utterances.

The results have shown that moderate talker identification and verification performance can be obtained in response to sequences of spoken digits of length 7 or greater. For random sequences, with duplication allowed, identification error rates range from 3.6 to 14.0 percent for talker populations ranging from 10 to 100 percent talkers. The average rank percentile of the correct talker is less than 0.8 percent for sequences of length 7 or greater, independent of population size. Thus, even when identification errors occur, the correct talker can be expected to rank quite high among all the candidates. For verification, average verification error rates (considered as the average of false reject and false accept rates) of 8 percent are obtained for sequence lengths 8 or greater. Reference profile adaptation is an important element in verification system performance, accounting for an improvement of 2 percent in error rate. In addition, for both identification and verification, the use of optimally weighted talker distances and two-class reference profiles each account for 1 or 2 percent improvements in error rates.

Performance variability among talkers remains a significant problem. Small groups of talkers account for persistently large identification and verification error rates. The tails in the distribution of individual talker error rates attributable to these talkers are particularly conspicuous for large sequence lengths, where the preponderance of individual talkers have attained quite low error rates.

Special problems arise in consideration of a basic premise of this sytem, that, within the context of the digit vocabulary, both the composition and length of a sequence of input utterances are optional. Much of the results have been devoted to describing the effects on performance of both the length and composition of digit sequences. Performance is quite poor for the input of single utterances, with error rates of the order of 18 percent, but improves rapidly with additional (distinct) utterances. Moderate error rates are obtained with the input of seven or more distinct utterances (at which point performance levels off). Repeated utterances of a digit provide little additional talker characterizing information and improvement in performance. For randomly selected digits, the sequence length should be one or two digits longer if repetitions are allowed to attain the same performance obtained for sequences with no repetitions. In addition, the absence of restrictions on the composition and length of input digit sequences complicates setting and adapting thresholds in the verification mode. Two procedures have been evaluated for setting and adapting thresholds based on the length and composition of the sequence. Procedure 2 has been shown to be reasonably reliable at maintaining a constant false reject rate as a function of sequence length. However, the procedures have been evaluated only for experiments in which the composition type and sequence length are the same for all trials.

To provide a frame of reference of the present system, the performance of some representative talker recognition systems can be cited. The performance of Doddington's talker verification entry control system, as cited in [6], is approximately 1 percent false accept rate with a false reject rate of 0.3 percent. Doddington's system is based on spectral matching in the vicinity of the vowel energy maximum of monosyllabic words. An input utterance consists of four words obtained with a broad-band high-quality recording in a sound booth. The system makes use of a "no decision" option in which verification decisions can be deferred until additional four-word phrases are obtained. The results quoted are obtained with an average of 1.3 phrases per talker.

Furui [7] reports average verification error rates of less than 1 percent with a system based on cepstral analysis of sentence-long utterances. The evaluation was carried out on recordings made in a sound booth over dialed-up telephone lines. In a subsequent, unreported evaluation of the same system carried out over a five-month period with talkers making verification trials from their own phones (usually in office environments), verification error rates of approximately 4 percent were obtained.

In contrast to these results obtained with techniques based on analysis of specified input utterance texts, systems based on text-independent analysis usually require considerably more input to attain comparable performance. For example, the best results obtained by Markel and Davis [8] for approximately 39 s of unconstrained voiced speech are 4 percent average verification error. Their system is based on long-term averages of reflection coefficients and fundamental frequencies over the voiced portions of random monologues obtained from high-quality telephone band recordings.

Each of these systems performs better than the system described in this paper, but each one requires considerably more computation and/or prototype storage, since they operate directly on the speech signal. The present system, with an average verification error rate of approximately 8 percent for eight-word utterances, operates only on the 12 distance scores per word provided by the word recognizer. Thus, although talker recognition performance for this sytem is not outstanding, considered in the perspective of its operation as a byproduct of talker-independent word recognition, it performs these tasks reasonably well with little additional cost in processing or storage other than what is required for the word recognition.

REFERENCES

[1] P. Calavrytinos, J. R. Wagner, and M. R. Schroeder, "Automatic word and speaker recognition based on time-frequency domain processing" (Abstr.), in *Proc. 9th Int. Congr. Acoust.*, Madrid, Spain, 1977, p. 485.

[2] R. L. Kashyap, "Speaker recognition from an unknown utterance and

speaker-speech interaction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 481-488, 1976.
[3] L. R. Rabiner, S. E. Levinson, A. E. Rosenberg, and J. G. Wilpon, "Speaker-independent recognition of isolated words using clustering techniques," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 336-349, 1979.
[4] R.O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*, New York: Wiley, 1973, pp. 230-235.
[5] A. E. Rosenberg and K. L. Shipley, "Speaker identification and verification combined with speaker independent word recognition," in *Proc. Int. Conf. Acoust., Speech, Signal Processing*, Apr. 1981, pp. 184-187.
[6] A. E. Rosenberg, "Automatic speaker verification: A review," *Proc. IEEE*, vol. 64, pp. 475-487, 1976.
[7] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-29, pp. 254-272, 1981.
[8] J. D. Markel and S. B. Davis, "Text-independent speaker recognition from a large linguistically unconstrained time-spaced database," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp.74-82, 1979.

Aaron E. Rosenberg (S'57-'63-SM'83-F'84) received the S.B. and S.M. degrees from the Massachusetts Institute of Technology, Cambridge, in 1960, and the Ph.D. degree from the University of Pennsylvania, Philadelphia, in 1964, all in electrical engineering.

He is a member of the Technical Staff in the Acoustics Research Department at AT&T Bell Laboratories, Murray Hill, NJ. He has been with Bell Labs since 1964, where his research interests have included auditory psychophysics, speech perception, and currently, speech and speaker recognition. He has authored or coauthored over 35 papers in these fields.

Dr. Rosenberg is a Fellow of the Acoustical Society of America and a member of Sigma Xi. He is a member of the Administrative Committee and the Conference Board of the IEEE Acoustics, Speech, and Signal Processing Society. He has also served as Chairman of the Society's Technical Committee on Speech Communication, and as Associate Editor for the IEEE TRANSACTIONS ON ACOUSTICS, SPEECH, AND SIGNAL PROCESSING.

Kathleen L. Shipley received the B.A. degree in mathematics from Douglass College, Rutgers University, New Brunswick, NJ, in 1970.

Since 1970 she has been a member of the Acoustics Research Department at AT&T Bell Laboratories, Murray Hill, NJ, where she is engaged in scientific programming for laboratory computer systems dedicated to research in communications acoustics.

Mrs. Shipley is a member of Pi Mu Epsilon.