Text-Independent Speaker Recognition from a Large Linguistically Unconstrained Time-Spaced Data Base

JOHN D. MARKEL, SENIOR MEMBER, IEEE, AND STEVEN B. DAVIS, MEMBER, IEEE

Abstract—A very large data base consisting of over 36 h of unconstrained extemporaneous speech, from 17 speakers, recorded over a period of more than three months, has been analyzed to determine the effectiveness of long-term average features for speaker recognition. Results are shown to be strongly dependent on the voiced speech averaging interval L_v. Monotonic increases in the probability of correct identification and monotonic decreases in the equal error probability for speaker verification were obtained as L_v increased, even with substantial time periods between successive sessions. For L_v corresponding to approximately 39 s of speech, text-independent results (no linguistic constraints embedded into the data base) of 98.05 percent for speaker identification and 4.25 percent for equal error speaker verification were obtained.

I. Introduction

In RECENT years, there has been an increasing interest in computer-based techniques for text-independent speaker recognition [1]-[6]. Recognition is used here to encompass both speaker identification and verification [7]. The term "text-independent" has been used in several different contexts. For example, Atal [1] has used the term in the sense of choosing independent randomized test frames from a single sentence to use against the remaining frames as a reference set. Sambur [4] has used the term in an experiment where the sentences in the test set were different from those in the reference set, even though each speaker read precisely the same list of sentences.

Although useful insight has been gained by these approaches, they were linguistically constrained. In many practical situations, where text-independent speaker recognition is desired, there typically will be no control over the speech being tested. As Beek, Neuberg, and Hodge [8] have pointed out, text-independent speaker identification can overcome problems which may arise if the speaker is uncooperative, and there is a great interest for speaker identification over communications channels, which have no linguistic constraints. Furthermore, there may be days to weeks of separation between reference and test sessions.

Several other studies [2], [3], [5], [6] have analyzed data with varying amounts of linguistic constraints. Li and Walker [2] used 30 s of speech read from the rainbow passage [9] recorded once by 22 male speakers and twice by an additional eight male speakers. They did not specify the number of days separating the recordings. They demonstrated that distances among spectral correlation matrices could be used to compare inter-speaker and intra-speaker differences. However, the same text was used for all tests, which could be interpreted as a linguistic constraint.

Hunt, Yates, and Bridle [6] used approximately six two- to three-min long FM radio weather forecasts from each of eleven male and two female speakers. Each forecast was divided into 20 or 30 s intervals and long-term fundamental frequency and cepstral coefficient features were computed for 20-ms sequential frames in each interval. They did not specify the number of days between successive forecasts by the same speaker. Using Fisher discriminant analysis [10], they achieved 89 percent correct speaker identification with independent test and reference sets. However, the speakers read text with some effort at uniformity between sessions, which could also be interpreted as a linguistic constraint.

In a preliminary study, Markel, Oshika, and Gray [5] used one 15- to 18-min interview from each of four male speakers with somewhat similar speech characteristics. The interviews were recorded with an audio tape recorder in a normal room environment. Long-term fundamental frequency, gain, and reflection coefficient features were computed for every 1000 sequential voiced frames (20-ms windows per frame, 50 frames/s) in each interview. Using the same Fisher discriminant analysis [10] as Hunt *et al.* to transform the data, they achieved perfect discrimination among the four speakers. These recorded interviews were considered to be free of linguistic constraints. However, the data were insufficient to obtain statistically significant results, and with only one session per speaker, there was no analysis of speaker characteristics over time.

The purpose of this paper is to present results from experiments in speaker recognition where there were no linguistic constraints on the speech content (other than the ones implied when the speaker is cooperative, and English is used). In comparison with the previous study [5], results are presented for a larger number of speakers, for multiple sessions from each speaker, and for a greater number of features. Furthermore, the effects of time between recording sessions are studied. For practical implementation, only parameters obtained from the analysis portion of a linear prediction vocoder (fundamental frequency, gain, and reflection coefficients) were used. (Beek

Manuscript received June 21, 1978; revised September 13, 1978. Part of this work was performed while the authors were at the Speech Communications Research Laboratory, Inc., Santa Barbara, CA 93109. This work was supported by the Advanced Research Projects Agency of the Department of Defense and was monitored by the Office of Naval Research under Contract N00014-77-C-0264.

The authors are with Signal Technology, Inc., Santa Barbara, CA 93101.

0096-3518/79/0200-0074$00.75

et al. [8] have stated that the reflection coefficients are currently favored for all-digital narrow-band communications systems.) This study shows that if these parameters are averaged over sufficiently long intervals of time, such as 30 s or more, the features obtained are essentially free of linguistic constraint, and speaker recognition performance is comparable with some text-dependent speaker recognition experiments. The linguistic results agree with Li and Walker [2], who used a smaller data base; long-term speech features are relatively stable after 30 s. Furthermore, this study shows that if the averaging interval is too short, speaker recognition performance is unacceptable with linguistically unconstrained extemporaneous speech. In addition, the importance of having a time-spaced reference set of sufficient size is demonstrated.

II. Data Base and Processing Methodology

A data base was collected by recording 170 15-min interviews from eleven male and six female speakers. There were ten sessions per speaker, with each session separated by a minimum of one week. Generally, the successive sessions were obtained within two to three weeks. One exceptional separation between successive sessions was fourteen weeks.

All sessions were recorded on a Tandberg 9000X two-track recorder at a recording speed of 7.5 in/s. One track was used to record the interviewer and the other track was used to record the speaker. The speaker was recorded with a B and K one-half-in condenser microphone and amplifier system in an IAC sound room equipped with a window. The interviewer was recorded with a conventional dynamic microphone outside of the sound room. Two-way communication was established using headphones.

Each session began with the speaker reciting his/her name, a password, a word list, and the first paragraph of the rainbow passage [9]. The interviewer posed a topic to the speaker, and the remaining time (generally twelve to thirteen min) was devoted to an extemporaneous monolog by the speaker. The interviewer responded briefly when appropriate, or when it was necessary to ask a new question for continuity.

A wide range of topics were covered, from describing a job to describing a frightening experience. Although one might argue that this approach in some sense constrained the data, casual listening of the recordings demonstrates that this is not the case. The topics generally provided a springboard for the speaker's thoughts, and the speech was usually conversational, fluent, and quite varied. (With one subject, the suggested topic was consistently replaced by a wide variety of topics.)

Several observations should be noted which may be of considerable importance in practical situations. After the initial recording gain calibration for each session, no further gain adjustments were made. Subjects occasionally became bored or distracted, and either lowered their voice intensity or turned their heads away from the microphone. Conversely, subjects occasionally became intense on a topic and nearly "swallowed" the microphone, resulting in substantial low-frequency waveform variability due to breath bursts. Also, there was some stuttering, throat clearing, laughter, giggling, and poor articulation.

In addition to these conditions, about half of the subjects acquired various degrees of colds during a two to three week period. All of these cases were recorded in the normal fashion, and no hand editing or deletion of any data was performed. The data used in this study consisted of only the extemporaneous speech material from the speakers, excluding the rainbow passage, word lists, etc. The total duration of the data base is 17 speakers X 10 sessions/speakers X approximately 13 min/session, or approximately 36.8 h of data.

Several large population and long duration data bases have been reported in the literature [10], [11]. These were all text-dependent studies with short names or phrases. However, even the total duration of the large data base used by Das and Mohn is only one-tenth the total duration of the data base used in this study. The magnitude of this data base was extremely valuable for choosing feature subsets and defining reference sets which spanned varying periods of time.

Each audio tape was manually cued to the location where the extemporaneous portion of the interview began. Then real-time linear prediction analysis and disk storage of the analysis parameters was initiated. The data were low-pass filtered at 3250 Hz and sampled at a 6500 Hz rate for compatibility with future applications to telephone systems and narrow-band vocoder systems. The speech samples were preemphasized with a factor of 0.9, successive 128-point frames were multiplied by a Hamming window, and the autocorrelation method of linear prediction was used at a rate of 50 frames/s. The analysis was performed in real time under Fortran control using a commercially available array processing system in conjunction with a PDP 11/45 computer [4], [12]. The analysis parameters for each speech frame were ten reflection coefficients, pitch period (obtained from a modified cepstral pitch tracker), and gain, and were stored in a quantized format of eight bits (one byte) per parameter. The process was terminated when the end of the tape was reached (defined as a 30-s silence interval). The processing of each interview resulted in an analysis file of approximately 1000 disk blocks (512 bytes/block), and all interviews together required nearly half the total space of a 200-Mbyte disk (340 670 formatted disk blocks). In comparison, it would require ten 200-Mbyte disks to digitize all of the interviews with 12 bits/sample and to store directly without preprocessing.

Next, the analysis files were used to obtain long-term feature vectors, where each vector was the average of L_v successive voiced analysis frames. Unvoiced and silence frames were not included in this study, since it was felt that fundamental frequency was an essential speaker-dependent parameter. The vocoder analysis parameters consisted of fundamental frequency (the reciprocal of the pitch period), gain, and ten reflection coefficients. For every interval L_v, long-term features based upon the mean, standard deviation, and dispersion (standard deviation divided by mean) of the twelve parameters were computed, resulting in 36-dimensional feature vectors. This feature set was defined in a reasonably general manner since analytic techniques for feature reduction may be used to find the most reasonable feature subsets for speaker recognition.

A summary of the number of feature vectors produced for all 170 interviews is given in Table I. In this table, the data are

TABLE I
NUMBER OF FEATURE VECTORS AND AVERAGE REAL-TIME INTERVAL (RTI) FOR EACH L_v CONDITION

		L_V			
		30	100	300	1000
REFERENCE SESSION	1-5	58,379	17,486	5,799	1,712
	6-10	58,032	17,376	5,764	1,701
TOTAL NUMBER OF TOKENS		116,411	34,862	11,563	3,413
AVERAGE SPEECH SEGMENT SIZE PER TOKEN (SEC)		1.14	3.80	11.47	38.85

partitioned into representative test and reference sets [13]. Four choices of L_v were studied, namely $L_v = 30$, 100, 300, and 1000. The total number of feature vectors and the average real-time interval per feature vector as functions of L_v are also given.

It is important to consider the relationship between a particular value of L_v and the real-time interval of a long-term feature vector. Most significantly, a fixed number of voiced frames, rather than all of the voiced frames from a fixed elapsed-time interval, was chosen for analysis. With extemporaneous speech, there may be intervals of 10 to 20 s where very little or no voiced speech occurs (the speaker may pause, cough, laugh, etc.), leading to a variable voiced frame rate. If long-term features were a function of the voiced frame rate, then such features would not be reflective of only a speaker's speech sounds, but also his/her speech rate and style. While these additional characteristics might be a source of speaker-dependent information, they were not considered in this study, and consequently long-term features were made independent of the voiced frame rate.

The real-time interval for a long-term feature (s/feature) corresponds to a product of the following factors: 1) the number of voiced frames per feature vector (L_v), 2) the reciprocal of the voiced frame to total frame ratio (or the reciprocal of the voicing duty factor), and 3) the reciprocal of the number of analysis frames/s (or the reciprocal of the frame rate). In a previous study [5], the voicing threshold was set such that very smooth fundamental frequency (F_0) contours were observed on a real-time display system, and as a result, $L_v = 1000$ corresponded to approximately 70 s of real speech. For this study, the voicing threshold was determined by synthesizing the speech using the F_0 contour obtained, and then selecting the threshold that produced the subjectively best synthesis. The ear appears more sensitive to voiced speech segments which are synthesized as unvoiced, rather than the reverse, i.e., buzziness is typically preferred over whispery or hoarse speech. As a result, more voiced decisions were made and $L_v = 1000$ in this study corresponded to approximately 39 s of speech.

The feature vectors for each interview for each of the above values of L_v required approximately 301, 93, 33, and 13 disk blocks, respectively, and a total of 74 800 disk blocks were required to store the feature vectors for the various L_v conditions for the 170 interviews. These data were then further processed as described in the next section.

III. EXPERIMENTS IN PARAMETER VARIABILITY

A. Intra-Speaker Variability

In a previous study [5], the within speaker (intra-speaker) variability of the features for one male speaker was demonstrated to be a monotonically decreasing function of L_v from $L_v = 1$ to $L_v = 1000$ for a single 15 min session. Using the data base in this study, it was possible to study the intra-speaker variability for a larger number of male and female speakers, and in addition, it was possible to study the intra-speaker variability for cumulative sessions. If individual sessions are described by $S(i)$, $i = 1, 10$, then cumulative sessions may be described by $C(i)$, $i = 1, 10$, where $C(i) = S(1) + S(2) + \cdots + S(i)$.

The standard deviations of the long-term averages of the fundamental frequency and the first reflection coefficient, denoted as $\sigma\langle F_0\rangle$ and $\sigma\langle k_1\rangle$, respectively, as measured over the cumulative sessions $C(i)$ for one male and one female speaker, are shown in Fig. 1.

For both speakers and for each set of cumulative sessions, $\sigma\langle F_0\rangle$ decreases as L_v increases. This behavior demonstrates that over long intervals, a speaker's average fundamental frequency is (probably) a good estimator of a characteristic or "habitual" value, and for successive long intervals, the deviation from the habitual value is small. For short intervals, influences such as speech prosody may mask the habitual value, and successive short intervals will deviate more widely from each other. This concept of habitual fundamental frequency is paralleled by the concept of habitual (perceived) pitch; the latter is used in speech therapy as a measure of acoustic improvement during treatment of a functional or organic voice disorder [14], and is an important factor in listener-based speaker recognition. For both speakers and for each value of L_v, there is a trend for $\sigma\langle F_0\rangle$ to increase as more sessions are included (although there are exceptions, e.g., for the female speaker, $\sigma\langle F_0\rangle$ for $C(1)$ is greater than $\sigma\langle F_0\rangle$ for $C(2)$). The dependence of $\sigma\langle F_0\rangle$ on L_v can approximately be described as proportional to $L_v^{-1/2}$, which agrees with the theoretical relationship between the variance of a set of samples of a stationary random process, e.g., the L_v samples of F_0, and the variance of the process [5]. In absolute terms, the standard deviation of the long-term fundamental frequency averages, over a time span of more than three months, varies from 17-23 Hz at $L_v = 30$ to 4-8 Hz at $L_v = 1000$ for the male speaker, and from 28-33 Hz at $L_v = 30$ to 8-11 Hz at $L_v = 1000$ for the female speaker.

The behavior of $\sigma\langle k_1\rangle$ as L_v increases mirrors the behavior of $\sigma\langle F_0\rangle$ as L_v increases. Since the value k_1 is a monotonic function of the spectral slope of a first-order linear prediction inverse filter for speech [5], [15], then a parallel explanation in terms of "habitual spectral slope" may be given, i.e., the longer the interval, the better the estimate of the habitual spectral slope. However, as more sessions are included, the behavior of $\sigma\langle k_1\rangle$ differs from the behavior of $\sigma\langle F_0\rangle$. For a given L_v, there is essentially no measurable increases in k_1 variability as the time period increases from one 15 min session to a period of nearly three months, with all ten sessions included. This trend is observed for the other speakers and the other long-term reflec-

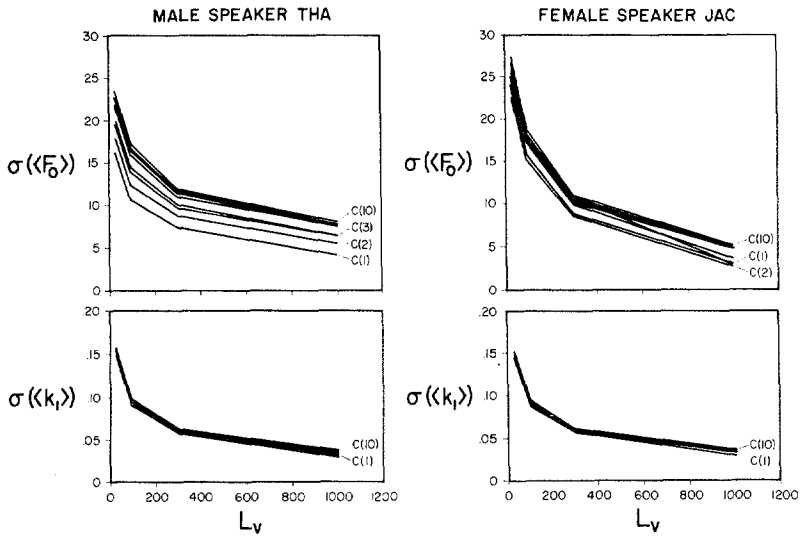

Fig. 1. Standard deviation of long-term features as a function of L_v, the number of voiced frames per feature vector.

tion coefficient averages, thus substantiating the presence of an "habitual spectral characteristic" for each speaker. Since the reflection coefficients are used to describe the vocal tract shape in an acoustic tube model [16], the result implies that the physical characteristics of a subject's vocal tract show no observable changes over at least several months.

Furui *et al.* [17]-[20] have examined speaker variability over intervals from a few weeks to several years. Their studies dealt with the variability of repeated word lists and short sentences. They found that for increasing time intervals from about three weeks to three months, spectral parameters such as reflection (PARCOR) or cepstral coefficients showed increasing variation. In contrast, the standard deviation of the reflection coefficients in this study show essentially no variation over time. Perhaps the data of Furui *et al.* were too linguistically constrained, and speakers never approached their habitual spectral characteristic.

In summary, inter-speaker variability based on averaged features decreases monotonically as the averaging interval increases. Furthermore, for a large averaging interval, inter-speaker feature variability is relatively consistent over a time period of three months. The next aspect of this study is a comparison which includes the intra-speaker information, e.g., a feature-by-feature analysis which uses the values of each feature from all subjects. If some features have small inter-speaker variance compared to the intra-speaker variance, then those features will not be useful for speaker recognition, and the performance of a classifier designed to recognize speakers from these features may be poor.

B. Variance Ratio Analysis

One method of measuring the usefulness of a feature for speaker recognition is the F-ratio or variance ratio (also referred to as the generalized Fisher discriminant) [7], [10], [19]. The variance ratio of a feature is the quotient of the inter-speaker variance and the intra-speaker variance [11]. In general, the larger the variance ratio for a particular feature, the greater the probable contribution of the feature in distinguishing the speakers [13], but this property is strongly dependent on the data and the experimental procedure. However, the variance ratio does not account for inter-feature correlations, and if two features with high variance ratios are highly correlated, then the inclusion of both parameters might be somewhat redundant [7].

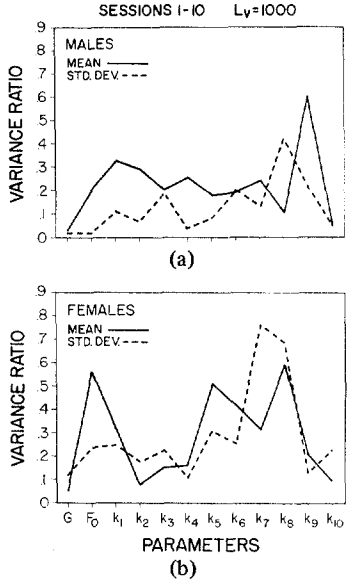

Fig. 2. Variance ratios from all 10 sessions as a function of long-term mean and standard deviations of parameters. (a) All male speakers. (b) All female speakers. $L_v = 1000$.

1) Trends as a Function of Population: The variance ratios for the case $L_v = 1000$ and cumulative sessions 1-10 are shown in Fig. 2 for the male and female speakers separately, and in Fig. 3 for two subsets of the male speakers. Only the variance ratios of the mean and standard deviation features are shown. The variance ratios of the dispersion features were consistently low, and therefore believed to contribute very little toward speaker recognition in this study.

There are noticeable differences in the variance ratios between the male and female populations. Based on relative magni-

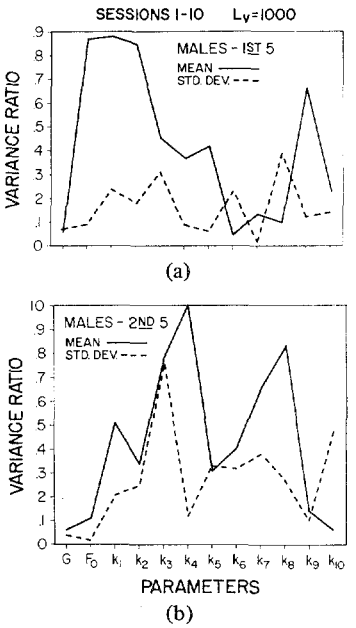

Fig. 3. Same conditions as Fig. 2 except: (a) Male speakers: First five; (b) Male speakers: Second five.

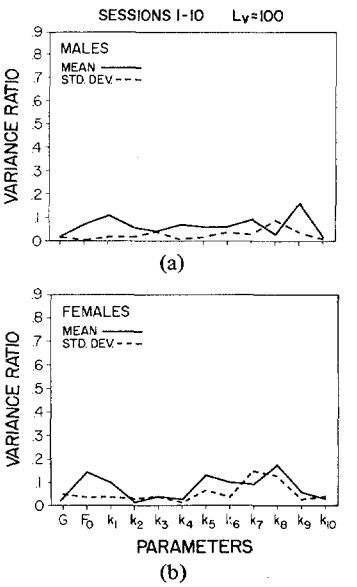

Fig. 4. Same conditions as Fig. 2 except that $L_v = 100$. (a) All male speakers. (b) All female speakers.

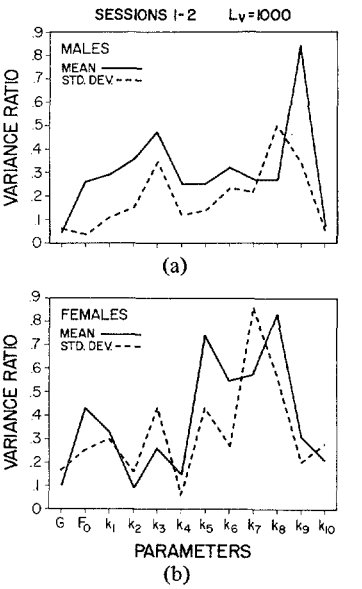

Fig. 5. Same conditions as Fig. 2 except only sessions 1-2 shown. (a) All male speakers. (b) All female speakers.

tudes, the features $\langle \sigma_{k_9} \rangle$, $\langle \sigma_{k_8} \rangle$, and $\langle k_1 \rangle$ would be the most significant for identifying the male population, while $\langle \sigma_{k_7} \rangle$, $\langle \sigma_{k_8} \rangle$, and $\langle k_8 \rangle$ would be the most significant for identifying the female population. If the male population is arbitrarily divided into two equal-sized subsets, there are pronounced changes in the variance ratios. For the first set of male speakers, $\langle k_1 \rangle$, $\langle F_0 \rangle$, and $\langle k_2 \rangle$ have the largest variance ratios, and for the second set of male speakers, $\langle k_4 \rangle$, $\langle k_8 \rangle$, and $\langle k_3 \rangle$ have the largest variance ratios. These results show the need to have a substantially larger speaker population in order to characterize the parameters of major importance. However, it is estimated that to obtain variance ratios which would exhibit consistent trends for a set of speakers and for subsets of the speakers, a much larger data base, possibly more than 100 speakers, would be required.

In the previous paper [5], for a smaller and more homogeneous data base, $\langle k_2 \rangle$ and $\langle k_6 \rangle$ were found to be the most significant parameters. These large variance ratios would be physical evidence for the importance of the first and third formants in voiced speech [5]. This larger population base, however, shows no such relationships. The conclusion is that for studies with linguistically unconstrained speech, parameter ranking using variance ratios should be used cautiously. The parameters with large variance ratios may change depending on how the data are partitioned, and the features with small variance ratios may be important for achieving good speaker recognition if the data partitioning is changed. (Conversely, it will be shown that some parameters with small variance ratios may actually *degrade* speaker recognition.)

2) Trends as a Function of L_v and Time-Spacing: The variance ratios were determined for the case $L_v = 100$ and cumulative sessions 1-10 (Fig. 4), and for the case $L_v = 1000$ and cumulative sessions 1-2 (Fig. 5). Comparing Figs. 2 and 4, which only differ by the averaging interval L_v, the variance ratios generally maintain the same relative relationships, i.e, the features which have the relatively larger variance ratios for $L_v = 1000$ also have the relatively larger variance ratios for $L_v = 100$. However, the absolute values of the variance ratios are smaller for $L_v = 100$ than for $L_v = 1000$. Comparing Figs. 2 and 5, which only differ by the number of sessions, the relative relationships and the absolute values of the variance ratios are similar for two cumulative sessions and for ten cumulative sessions. However, a slight decrease in the absolute values of the variance ratios for ten cumulative sessions is observed. If the inter-speaker variance is assumed to be relatively constant for two or ten cumulative sessions, then the slight decrease in variance ratios for ten sessions over two sessions correlates with the slight increase in standard deviations observed in Fig. 1. This result further establishes that a speaker's habitual features, when measured over a relatively long interval (greater

than 30 s), do not show appreciable changes over time periods up to three months.

3) Further Observations: It is also evident that the variance ratios for the mean features generally have larger values than the corresponding variance ratios for the standard deviation features. The variance ratios for the dispersion features are in turn substantially lower in value than the corresponding variance ratios for the standard deviation features. Features based upon gain have consistently small variance ratios.

IV. Speaker Recognition

Speaker recognition was based on a weighted Euclidean distance metric [5], [7], [11], where the mean vector and inverse covariance matrix for each of the 17 speakers were estimated from feature vectors in the specified reference set. All 36 dimensions were used initially. The distances between each reference class and each test vector were computed, and the test vector was assigned to the reference class which yielded the smallest distance. For speaker identification, a tally was taken of the number of correct choices. For speaker verification, the distances were stored for further analysis with a variable distance threshold. The method of cross-validation in both directions was used [11], where independent subsets of the data were cyclically treated as test and reference groups, and the speaker recognition scores for each cycle were averaged for the final scores.

Atal [7] and Bricker *et al.* [10] discussed three possible choices for a distance metric. Each metric was a positive semi-definite form which could be described by $d = (X - Y_i) \cdot M(X - Y_i)^T$, where X was a vector to be classified, Y_i was the mean vector for class i, and M was a weighting matrix. The choices for M were a pooled covariance matrix W^{-1} from all speakers, an individual covariance matrix W_i^{-1} from each speaker, or a discriminant matrix D composed of the eigenvectors of $W^{-1}B$, where B was the between-class covariance matrix.

The use of the discriminant matrix D requires sufficient knowledge of the inter-speaker variability, which may be difficult to attain unless an extremely large number of speakers is used. Atal [7] and Bricker *et al.* [10] preferred the pooled covariance matrix W^{-1} over the individual covariance matrix W_i^{-1}. Their rationale was that data limitations (less samples than dimensions) frequently result in a singular (noninvertible) covariance matrix, and that one pooled covariance matrix would adequately represent all speakers, even though speaker dependent data is contained in individual covariance matrices and subsequently is not used.

From Table I, the average number of feature vectors per speaker per session for $L_v = 30$, 100, 300, 1000 is 685 (116 411/170), 205 (34 862/170), 68 (11 563/170), and 20 (3413/170), respectively. For $L_v = 30$, 100, or 300, with 36 dimensions, the individual covariance matrices were never singular for any number of pooled sessions. For $L_v = 1000$, with 36 dimensions, the individual covariance matrices were singular if less than three sessions are pooled. Furthermore, Kanal [14] has suggested that ten times the number of dimensions is an adequate sample size for good covariance matrix estimates with normal probability distribution assumptions. For five sessions and 36 dimensions in a reference class, the factors which relate sample size to dimensionality for $L_v = 30$, 100, 300, 1000 are 95 (685 · 5/36), 28 (205 · 5/36), 9 (68 · 5/36), and 3 (20 · 5/36), respectively. For $L_v = 1000$, sessions as long as 45 min would have been necessary to produce a factor near ten, but a factor as large as ten is probably not needed for features which are themselves the average of 1000 frames of data. However, 15 min was a sufficient duration for the other values of L_v, as well as an upper limit of endurance for the subject and interviewer. It was felt that the advantages gained through the use of individual covariance matrices outweighed potential problems of undersampling the speaker's statistics. In a practical situation, relatively long sessions would be necessary for sufficient accumulation of speaker's reference data, but thereafter the speaker could be verified approximately every 39 s.

TABLE II
Speaker Recognition Based on Partitioning Data in Half and with 36 Long-Term Features

SPEAKER IDENTIFICATION
Percent of correct choices based on minimum distance

SESSION		L_v			
REF.	TEST	30	100	300	1000
1-5	6-10	61.20	78.65	88.20	93.34
6-10	1-5	59.87	75.48	85.27	89.77
AVERAGE		60.54	77.06	86.74	91.56

A

SPEAKER VERIFICATION
Percent of false acceptances and false rejections based on equal error criterion

SESSION		L_v			
REF.	TEST	30	100	300	1000
1-5	6-10	43.4	27.8	10.7	9.4
6-10	1-5	42.8	26.9	10.5	8.2
AVERAGE		43.1	27.4	10.6	8.8

B

SPEAKER VERIFICATION
Threshold distance based on equal error criterion

SESSION		L_v			
REF.	TEST	30	100	300	1000
1-5	6-10	5.79	7.52	9.78	18.84
6-10	1-5	5.85	7.58	10.85	21.10
AVERAGE		5.82	7.55	10.32	19.97

C

A. Trends as a Function of L_v

For the first series of tests, the first five sessions were treated as the reference data, the second five sessions were treated as the test data, and then vice versa. Results are shown in Table II. In Table II-A, it is seen that the average scores for the probability of correct identification P(CI) monotonically increase from 60 percent to nearly 92 percent as L_v increases from 30 to 1000, respectively. A confusion matrix of identification errors shows that no one speaker is more difficult to identify than any other speaker. In Table II-B, as L_v increases, the speaker verification equal error probability [probability of false acceptance P(FA) equals the probability of false rejection P(FR)] monotonically decreases from 43.1 percent to 8.8 percent. This trend is principally due to the P(FA) behavior, since the P(FR) behavior does not change appreciably with L_v [5]. Although the distance threshold for a given probability of correct acceptance and fixed dimensionality (under multivariate normal assumptions) may be analytically obtained, the

distance threshold for the equal error probability can only be determined experimentally. In Table II-C, the equal error probability distance threshold is seen to monotonically increase as L_v increases.

It is interesting to illustrate the difference between text-independent speaker recognition with and without linguistic constraints. Sambur has proposed an orthogonal linear prediction set of parameters for test-independent speaker recognition [4]. Within the context of a linguistically constrained experiment where all speakers spoke the same set of sentences, Sambur's text-independent results (in the sense that the reference sentences were different from the test sentences) were near 94 percent. The orthogonal linear prediction parameters are essentially equivalent to a linear transformation of the long-term reflection coefficients averages used in this study if $L_v = 1$ (equivalent to no averaging). If all linguistic constraints are removed, and if little or no averaging is used, the results of Table II indicate that the speaker identification scores for a true text-independent situation with a reasonable number of speakers will be quite poor (even for $L_v = 30$, P(CI) is bounded from above at 62 percent). A similar statement follows for the case of speaker verification.

B. Trends as a Function of Time Spacing

Rosenberg [21] has noted that one of the most important considerations in designing a data base is the time period over which utterances are collected and the methods for establishing reference patterns over time. Following the pictorial scheme of Furui *et al.* [14]-[17] for illustrating reference and test sets over time, speaker recognition for four cases shown in Fig. 6 were investigated. Reference sets were composed of from two to five succession sessions (with a time interval of at least one week between sessions). No co-mingling such as odd-numbered reference sessions and even-numbered test sessions was allowed. For each case, the reference and test sets were composed of equal numbers of successive independent sessions, and two-direction recognition tests (as described above) were made for the four L_v cases.

The results are presented in Table III. It is seen that for all L_v conditions, higher scores were obtained as the number of cumulative sessions increased.

The differences in the speaker identification score between the first two sessions and the first five sessions is around 15 percent for all L_v cases shown ($L_v = 1000$ was not used for two sessions since the covariance matrices were singular). It is interesting to note that in a text-dependent speaker verification experiment with different parameters and approaches, Luck [22] found that speech samples collected over a five week period gave the best results.

C. Trends as a Function of Feature Subsets

In a previous section, it was noted that the dispersion features had very small variance ratios, whereas the mean features as a group consistently had the largest variance ratios. How would recognition scores compare if the dispersion features were omitted, or if only the mean features were included? The recognition test for $L_v = 1000$ and five sessions per reference and test set was repeated using several different feature subsets, based on an analysis of the magnitudes of the variance

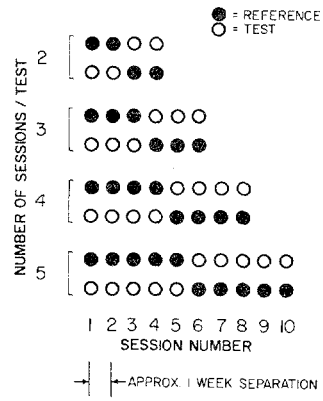

Fig. 6. Relations between reference samples and test samples for experimental results of Table III.

TABLE III
PERCENT OF SPEAKERS CORRECTLY IDENTIFIED AS A FUNCTION OF THE NUMBER OF REFERENCE SESSIONS

SESSIONS			L_v			
NO.	REF.	TEST	30	100	300	1000
2	1-2	3-4	50.36	64.34	71.18	—
2	3-4	1-2	53.45	67.95	75.31	—
3	1-3	4-6	54.29	70.03	79.12	80.58
3	4-6	1-3	57.04	72.69	82.14	89.30
4	1-4	5-8	59.91	76.41	86.73	92.85
4	5-8	1-4	59.26	74.62	83.45	86.34
5	1-5	6-10	61.20	78.65	88.20	93.34
5	6-10	1-5	59.87	75.48	85.27	89.77

ratios. In one case, only the twelve mean features were used, and in a second case, only the 24 mean and standard deviation features were used. The average scores for the two cases were P(CI) = 93.6 percent with P(FA) = P(FR) = 14.5 percent, and P(CI) = 96.8 percent with P(FA) = P(FR) = 7.2 percent, respectively. For comparison, the comparable average scores for all 36 features (Table II) were P(CI) = 91.6 percent with P(FA) = P(FR) = 8.8 percent.

Not only did both of these new cases based on feature subsets yield better scores than the original 36 dimension feature set, but in the second case, the identification score was markedly increased by more than 5 percent. This result is a significant practical illustration that the inclusion of some parameters which would hopefully improve performance (or at the worst case would have no effect on performance), can sometimes actually degrade the system performance in an open test. In a closed test with the distance metric used in this study, where a reference set also is used as a test set, this theoretically cannot happen. Closed tests on this data base verified that monotonic increases in the number of features produced monotonic increases in the P(CI) and monotonic decreases in equal error probability P(FA) = P(FR).

This improved performance by eliminating features with relatively small variance ratios was the basis for one additional test with a feature subset. In considering the remaining 24 features, the gain-related features had very small variance ratios, and furthermore, the inclusion of gain-related features was difficult to physically justify. In fact, it could be argued that even if these features helped, they should not be included because they may simply reflect a speaker's position, interest, etc. during the recording session. Therefore, the recognition test with only 24 features was repeated with the gain-related

TABLE IV
PERFORMANCE WITH FUNDAMENTAL FREQUENCY AND REFLECTION COEFFICIENT MEAN AND STANDARD DEVIATION LONG-TERM FEATURES, $L_v = 1000$ (AVERAGE REAL-TIME INTERVAL = 39 s)

FINAL RESULTS OF 2-WAY TESTING ON 38 HOURS OF EXTEMPORANEOUS SPEECH

SESSION NUMBERS		SPEAKER IDENTIFICATION	SPEAKER VERIFICATION
REF.	TEST	P(CI) (%)	P(FA) = P(FR) (%)
1-5	6-10	98.65	3.3
6-10	1-5	97.45	5.2
AVERAGE SCORE		98.05	4.25

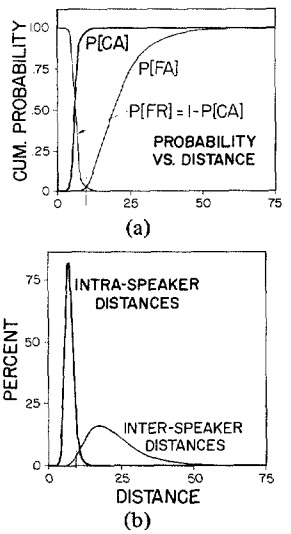

Fig. 7. Intra- and inter-speaker comparisons. (a) Cumulative probability. (b) Probability density estimates.

features removed, and the performance of this last with only 22 parameters was better than any previous test. The final results of this study using only the 22 fundamental frequency and reflection coefficients long-term averages are shown in Table IV. These results are extremely promising for future studies in many areas of speaker recognition. This substantially large testing effort (over eighty million distance measurements) has shown that realistic and acceptable speaker identification and speaker verification can be achieved with text-independent linguistically unconstrained speech.

The cumulative probability functions [Fig. 7(a)] and the probability density functions [Fig. 7(b)] for false rejection and false acceptance may be used to compare the inter- and intra-speaker distances in the verification task. These curves are derived from the first half of the final speaker verification test with 22 features, $L_v = 1000$, reference sessions 1-5 and test sessions 6-10. The equal error point is graphically depicted as the crossover point of the two cumulative probability curves in Fig. 7(a). This equal error point is found at a distance threshold where the probability of false acceptance (i.e., acceptance of an imposter) is equal to the probability of false rejection (i.e., rejection of a correct speaker).

The probability density functions (pdf's) in Fig. 7(b) show the distribution of the intra- and inter-speaker distances. The crossover point in Fig. 7(a) divides each of the pdf's into two sections, with the area under the intra-speaker pdf to the right

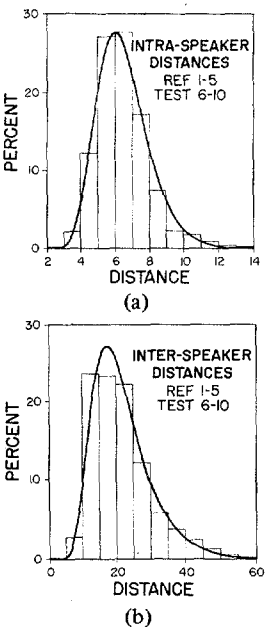

Fig. 8. Distance histograms and models. (a) Intra-speaker distances. (b) Inter-speaker distances.

of the dividing line equal to the area under the inter-speaker pdf to the left of the dividing line. For this data, the equal error crossover point is close to the intersection of the two pdf's, but only identical and symmetric pdf's will always have identical crossover and intersection points.

For test sessions 6-10 with $L_v = 1000$, there were a total of 1708 test vectors from the 17 speakers. The distances between each of thest test vectors and the correct reference speaker comprise the intra-speaker distance space. A histogram of these intra-speaker distances is shown in Fig. 8(a). The mean and standard deviation of the histogram distances were used to approximate normal and log-normal distributions. For the open test, there is no underlying theoretical distribution, and a chi-square test was used to measure the goodness of fit of the normal and log-normal distributions. The log-normal distribution had the smallest chi-square measure. Analogously, the distances between each of the 1708 test vectors and each of the 16 incorrect speakers (i.e., eliminating the reference speaker who is a correct match to the text vector) comprise the inter-speaker distance space. A histogram of the 27 318 inter-speaker distances is shown in Fig. 8(b). A log-normal distribution is a better fit to the inter-speaker histogram than a normal distribution, but not as good a fit as with the intra-speaker histogram.

V. SUMMARY

The significance and value of long-term feature averaging for text-independent speaker recognition with linguistically unconstrained speech has been demonstrated. This study used practical analysis conditions of telephone-range spectral width (0-3250 Hz) and parameters obtained from a linear prediction vocoder. All parameter-related computations were performed in real time using 16-bit integer arithmetic, and all parameters were further quantized into an 8-bit format for efficient disk storage.

The recording environment was controlled by recording the speakers with a condenser microphone in an IAC sound room.

An important extension of this work would be to reprocess the "clean-text" audio tapes through various channel disturbances such as the telephone system to determine the robustness of the approach in less ideal environmental conditions [20]. Also, in some situations, reference data may be obtained in a clean environment and subsequent speaker recognition attempted in a noisy environment. This area also requires investigation.

Although 17 speakers is not a trivial population size, it appears that for determining the importance of individual features for speaker recognition using linguistically unconstrained text, a substantially larger population base is required. It was found that features obtained from only one or two sessions of a given population are relatively unchanged over a much larger number of time-spaced sessions, where there was at least one week between sessions. Other features should also be investigated. It has been suggested that mean deviations [21] may prove more useful than the standard deviations used in this study. Further research is also required to assess the conditions, e.g., the number of long-term samples from a speaker, for obtaining a good estimate of the mean and variance of a speaker's characteristics.

An assumption throughout has been that only voiced speech frames are to be used in the analysis. If this assumption was not necessary, or if only slight degradation occurred if both voiced and unvoiced speech frames were included, the process would be simplified computationally, and in addition, 1000 frames per average would correspond to a real-time interval only about half as long as required here.

The best speaker recognition was obtained when 1) five sessions successively separated by at least one week were used to define the reference set, 2) the mean and standard deviation of the long-term averages of the fundamental frequency and reflection coefficients were used, and 3) each feature was obtained by averaging 1000 voiced analysis frames (corresponding to average real-time intervals of about 39 s). With approximately 18 h of reference data and 18 h of independent test data from 17 speakers, spaced over nearly three months in time, an average speaker identification score of 98.05 percent and an average equal error speaker verification rate of 4.25 percent were measured.

Acknowledgment

The authors would like to acknowledge the substantial contributions of the following people on this project: Dr. B. Oshika for organization of the data base, and R. Arnott and T. Applebaum for programming assistance with the real-time system and the nonreal-time batch processing programs. Also, the authors appreciate the helpful comments from Dr. J. Wolf and Dr. M. Hunt on an early draft of this paper.

References

[1] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech waves for automatic speaker identification and verification," *J. Acoust. Soc. Amer.*, vol. 55, pp. 1304–1312, 1974.

[2] K. P. Li and G. W. Walker, "Talker differences as they appear in correlation matrices of continuous speech spectra," *J. Acoust. Soc. Amer.*, vol. 55, pp. 833–837, 1974.

[3] K. O. Mead, "Identification of speakers from fundamental frequency contours in conversational speech," *Joint Speech Res. Unit*, Rep. 1002, 1974.

[4] M. R. Sambur, "Speaker recognition using orthogonal linear prediction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 283–289, Aug. 1976.

[5] J. D. Markel, B. T. Oshika, and A. H. Gray, Jr., "Long-term feature averaging for speaker recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-25, pp. 330–337, Aug. 1977.

[6] M. J. Hunt, J. N. Yates, and J. S. Bridle, "Automatic speaker recognition for use of communication channels," in *Conf. Rec., 1977 IEEE Int. Conf. Acoust., Speech, Signal Processing*, p. 764.

[7] B. S. Atal, "Automatic recognition of speakers from their voices," *Proc. IEEE*, vol. 64, pp. 460–475, Apr. 1976.

[8] B. Beek, E. P. Neuberg, and D. C. Hodge, "An assessment of the technology of automatic speech recognition for military applications," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-25, pp. 310–322, Aug. 1977.

[9] G. Fairbanks, *Voice and Articulation Handbook*. New York: Harper, 1960.

[10] P. D. Bricker, R. Gnanadesikan, M. V. Mathews, S. Pruzansky, P. A. Tukey, K. W. Wachter, and J. L. Warner, "Statistical techniques for talker identification," *Bell Syst. Tech. J.*, vol. 50, pp. 1427–1454, 1971.

[11] S. K. Das and W. S. Mohn, "A scheme for speech processing in automatic speaker verification," *IEEE Trans. Audio Electroacoust.*, vol. AU-19, pp. 32–43, Mar. 1971.

[12] R. D. Arnott and J. D. Markel, "Fortran control of real-time signal processing with high-speed processors," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-26, pp. 278–284, Aug. 1978.

[13] L. Kanal, "Patterns in pattern recognition: 1968–1974," *IEEE Trans. Inform. Theory*, vol. IT-20, pp. 697–722, Nov. 1974.

[14] L. Travis, *Handbook of Speech Pathology and Audiology*. New York: Appleton, 1971.

[15] A. H. Gray, Jr. and J. D. Markel, "A spectral-flatness measure for studying the autocorrelation method of linear prediction of speech analysis," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-22, pp. 207–217, June 1974.

[16] H. Wakita, "Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms," *IEEE Trans. Audio Electroacoust.*, vol. AU-21, pp. 417–427, Oct. 1973.

[17] S. Furui and F. Itakura, "Talker recognition by statistical features of speech," *Electron. Commun. Japan*, vol. 56-A, pp. 62–71, 1973.

[18] S. Furui, F. Itakura, and S. Saito, "Talker recognition by long-time averaged speech spectrum," *Electron. Commun. Japan*, vol. 55-A, pp. 54–61, Oct. 1972.

[19] S. Furui, "An analysis of long-term variations of feature parameters of speech and its application to talker recognition," *Electron. Commun. Japan*, vol. 57-A, pp. 34–42, 1974.

[20] S. Furui, F. Itakura, and S. Saito, "Personal information in the long-time averaged speech spectrum," *Review Elec. Commun. Lab.*, vol. 23, pp. 1133–1141, 1975.

[21] A. E. Rosenberg, "Automatic speaker verification: A review," *Proc. IEEE*, vol. 64, pp. 475–487, Apr. 1976.

[22] J. E. Luck, "Automatic speaker verification using cepstral measurements," *J. Acoust. Soc. Amer.*, vol. 46, pp. 1026–1031, 1969.