and nonstationary formulations of linear prediction applied to voiced speech analysis," presented at the 87th Meeting of the Acoustical Society of America, New York, NY, Apr. 1974 (A).
[11] F. Itakura and S. Saito, "Speech analysis-synthesis system based on the partial autocorrelation coefficient," presented at the Acoustical Society of Japan Meeting, Oct. 1969.
[12] J. R. Haskew *et al.*, "Results of a study of the linear prediction vocoder," *IEEE Trans. Commun.*, vol. COM-21, pp. 1008–1015, Sept. 1973.
[13] J. D. Markel and A. H. Gray, Jr., "A linear prediction vocoder simulation based upon the autocorrelation method," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-22, pp. 124–134, Apr. 1974.
[14] W. C. Lin and A. Agrawal, "Minicomputer-based laboratory for speech intelligibility research," *Proc. IEEE Conf. on Minicomput.*, Nov. 1973.
[15] J. N. Holmes, "Variable-frame-rate coding scheme for speech-analysis/synthesis systems," *Electron. Lett.*, Apr. 1974.
[16] J. L. Flanagan, *Speech Analysis Synthesis and Perception*, 2nd ed. New York: Springer-Verlag, 1972.
[17] S. Chandra and W. C. Lin, "Updating of speech transmission parameters obtained from nonuniform analysis segments in a linear predictive speech compression system," presented at the 87th Meeting of the Acoustical Society of America, Apr. 1974 (A).
[18] J. D. Markel and A. H. Gray, Jr., "On autocorrelation equations as applied to speech analysis," *IEEE Trans. Audio Electroacoust.*, vol. AU-21, pp. 69–79, Apr. 1973.
[19] J. D. Markel, "The SIFT algorithm for fundamental frequency estimation," *IEEE Trans. Audio Electroacoust.*, pp. 367–378, Dec. 1972.
[20] S. Chandra, "Linear prediction speech compression system and its evaluation," Ph.D. dissertation, Dep. Comput. Inform. Sci., Case Western Reserve Univ., Cleveland, OH, June 1974.
[21] S. Seneff, "A new encoding technique for the *k*-parameter: A statistical approach," M.I.T. Lincoln Lab., Lexington, MA, NSC Note 53.
[22] F. Itakura and S. Saito, "On the optimum quantization of feature parameters in the PARCOR speech synthesizer," in *Proc. 1972 Conf. on Speech Communication and Processing*, pp. 434–437.

Long-Term Feature Averaging for Speaker Recognition

JOHN D. MARKEL, MEMBER, IEEE, BEATRICE T. OSHIKA, AND AUGUSTINE H. GRAY, JR., MEMBER, IEEE

Abstract—**The potential benefits of long-term parameter averaging for speaker recognition were investigated. Parameters studied were pitch, gain, and reflection coefficients. Parameter variability was computed over various averaging lengths from one frame averaging (in effect, no averaging) to 1000 frame averaging (about 70 s of speech). It was demonstrated that the between-to-within speaker variance ratio, measured over several speakers, was significantly increased by performing long-term averaging of the parameter sets. The reflection coefficient averages for k_2 and k_6, respectively, were shown to produce the highest variance ratios.**

I. Introduction

THERE have been several studies on the choice of acoustic features in speaker recognition tasks [14], [19], [22]. Average fundamental frequency has been found to be a useful discriminating feature [13], as have gain measurements [2], [10] and long-term speech spectra [4]-[6], [9]. Perceptual studies indicate that "there is at least weak evidence that a voice that is distinctive to listen to also has distinctive spectrographic patterns" [20], and that dimensions of "characteristic pitch" and "characteristic loudness" may be posited to differentiate among speakers [21]. These speaker characteristics can be distinguished from the acoustic cues which signal linguistic elements, e.g., phonemes or words. For example, the realization of the word "bit" by a female child is acoustically very different from the same word pronounced by an adult male, yet the words are generally understood to be equivalent while the speakers are clearly different. It appears, then, that listeners adapt to speakers' *voice* characteristics (as well as their *linguistic* characteristics).

All this suggests that there are long-term characteristics which can be used in text-independent speaker recognition tasks. Such characteristics include long-term averages related to fundamental frequency, gain, and spectral averages.

The motivation for long-term averaging in text-independent speaker recognition is based upon a result from statistical sampling theory.

We assume that $\{p(i)\}$ defines statistically independent, identically distributed samples of the parameter p with true mean μ_p and variance σ_p^2. (For example, $\{k_1(i)\}$ corresponds to the reflection coefficient k_1 samples for each analysis frame.) If $x = \langle p(i) \rangle$ defines a *feature* based upon long-term averaging of p, where

$$\langle p(i) \rangle = \frac{1}{L_v} \sum_{i=0}^{L_v-1} p(i), \tag{1}$$

and L_v is the number of voiced analysis frames used in the averaging, then the variance of x is given in terms of the original

Manuscript received May 13, 1976; revised December 13, 1976 and March 15, 1977. This research was supported by the Advanced Research Projects Agency of the Department of Defense and was monitored by the Office of Naval Research under Contract N00014-73-C-0221.

J. D. Markel and B. T. Oshika are with the Speech Communications Research Laboratory, Inc., Santa Barbara, CA 93109.

A. H. Gray, Jr., is with the Speech Communications Research Laboratory, Inc., Santa Barbara, CA 93109 and the Department of Electrical Engineering and Computer Science, University of California, Santa Barbara, CA 93109.

parameter variance σ_p^2 by

$$\sigma^2[\langle p(i)\rangle] = \sigma_p^2/L_v. \quad (2)$$

The sample variance as a function of L_v is an important figure of merit for a particular feature. For example, if the features are more tightly clustered together about the sample mean as L_v increases from $L_v = 1$ (no averaging), then the intraspeaker variability is decreased, and the parameters would be expected to result in higher performance in text-independent speaker recognition tasks. Although no "true mean or true variance" exists for real speech because of physiological variations in human speech, it is reasonable to assume that at least some convergence or clustering of parameters will occur with long-term averaging.

The purpose of this paper is to define several sets of potentially useful long-term features and then to investigate their statistical properties as a function of the averaging length L_v. In addition, discrimination tests are presented over a small homogeneous set of speakers to illustrate the potential benefits of long-term averaging for unconstrained text-independent speaker recognition.

II. Features

To discuss the applicability of long-term feature averaging in a quantitative manner, we have chosen three different feature sets as the basis for analysis. Some of these features reflect physiological characteristics more closely than others.

A. Fundamental Frequency Features

Due to physiological considerations such as the length and thickness of the vocal folds, and respiratory muscle patterns, the phonation of a particular vowel with "normal effort" may result in differing rates of vocal fold vibration (corresponding to the acoustical correlate of fundamental frequency) for different speakers. For example, a child will have a high fundamental frequency compared to an adult because of the child's smaller vocal folds.

Although fundamental frequency, along with intensity and duration, is a controllable attribute of stress and intonation which may vary widely, each person appears to have a mean fundamental frequency value which, if averaged over a sufficiently long period of time, is relatively constant over a reasonable time span and is independent of linguistic content [8].

In addition, the standard deviation of the fundamental frequency over a long interval of time may carry important speaker-dependent information. For example, if the speaker is judged to be a monotone speaker, then the standard deviation would be expected to be relatively small. However, if the speaker is thought to be an "expressive" or "forceful" speaker, it would be expected to be relatively large.

B. A Gain Feature

It seems reasonable to assume that one of the characteristics that contributes to a speaker's identity is the amount of intensity or gain variation in his speech over time. Subjectively, the amount of gain variation is possibly correlated with the perception of "dynamic" versus "flat" voices. The actual gain variation is also a function of phonetic content, word and phrase stress, and discourse context. For example, for a constant subglottal pressure, the acoustical output energy for an /a/ is about 5 dB greater than for a /u/. Also, a larger gain variation would be expected with an exclamatory as opposed to a normal declarative sentence. Our assumption is that, over a sufficiently long interval of speech, gain variation can be considered part of the individual speaker's characteristics. That is, a speaker who is judged overall to be an "emphatic" speaker will have larger gain variation than one who is judged to have a usually monotonous voice.

In the measurement of gain variation, it is very important that results be only a function of speaker characteristics and not absolute system gain. Furthermore, because of the distinctly different production mechanism between voiced and unvoiced speech, it is desirable to measure the gain variations only during voiced speech. A normalized gain variation which satisfies desired physical properties is now defined. If $R(n)$ defines the energy of N speech samples $\{s(l)\}$ in frame n, then

$$R(n) = \sum_{l=0}^{N-1} s^2(l). \quad (3)$$

The sample mean and sample variance of $R(n)$ over L_v voiced frames is then defined by

$$\bar{R} = \langle R(n)\rangle \quad (4)$$

and

$$\sigma_R^2 = \langle (R(n) - \bar{R})^2\rangle \quad (5)$$

where $\langle\cdot\rangle$ will be used throughout to denote averaging over L_v voiced frames. The normalized gain variation δ is then defined by

$$\delta = \sigma_R/\bar{R}. \quad (6)$$

If the overall system gain is changed by a constant value, δ is unaffected. Furthermore, δ is nonnegative with $\delta = 0$ only when $\sigma_R = 0$. Physically, $\delta = 0$ means that the speech envelope (more precisely, the frame energy) is unchanged over the complete range of voiced speech analyzed.

C. Spectral Features

It is well established in the literature that one of the acoustical features that tends to differentiate one particular speaker from another during voiced speech production is the glottal sound source shape [15].

Although the spectral slope of a single glottal pulse can vary over a wide range from nearly whispered speech to very intense vocal effort, for normal conversational speech it is expected that an average glottal source spectrum could be obtained over a relatively long interval of speech that would have relatively small intraspeaker variability.

Unfortunately, glottal volume velocity waveform estimation from speech is a nontrivial task [7], [12], [16]. A more direct method for automatic real-time analysis is to use a parameter set that is related to the smooth characteristics of the spectrum, which is independent of fundamental frequency or gain. With linear prediction analysis, obvious possibilities are filter coefficients, reflection coefficients, or log area functions. Sambur [17] compared these coefficients in a speech recogni-

tion experiment and decided to make use of the reflection coefficients. Although reflection coefficients are nonlinearly related to the more physically meaningful smooth-spectral and log-spectral model from linear prediction analysis, there is ample evidence that they do contain important speaker-dependent information that is not contained in fundamental frequency- or gain-related parameters. For example, in the case of a first-order filter, $M = 1$, a smooth spectral model can be physically and mathematically related to the first reflection coefficient. This model [11, p. 139] has a spectral flatness given by

$$\Xi(1/A) = 1 - k_1^2. \tag{7}$$

If the speech sample being analyzed has a nearly flat spectral trend, k_1 approaches zero and the spectral flatness approaches unity. As the spectral slope increases negatively, k_1 approaches -1 and the spectral flatness approaches zero.

Based upon the spectral matching properties of linear prediction [11, p. 134], we would assume that preemphasis of the data would be beneficial since the reflection coefficients would then carry more information about the spectral structure at higher frequencies.

It would also seem reasonable that if long-term, spectrally related features are desired which minimize intravariability, only voiced speech should be analyzed. Substantial differences exist in the physiological mechanisms which produce voiced and unvoiced sounds. Since the excitation for unvoiced speech is generally assumed to have a flat spectrum, the difference in spectral slope between voiced and unvoiced sounds may be on the order of 8-16 dB. With only voiced sounds, some variation will still occur since different articulator positions will cause variations on the acoustic loading at the glottis, affecting the glottal source shape. This variation, however, is expected to be substantially less than that due to glottal source variations in voiced-unvoiced speech production.

D. Summary of Feature Definitions

As *features* we study the following.

1) F_0 average

$$x_1 = \overline{F_0} = \langle F_0(n) \rangle. \tag{8}$$

2) Standard deviation of F_0

$$x_2 = \sigma_{F_0} = \langle [F_0(n) - \overline{F_0}]^2 \rangle^{1/2}. \tag{9}$$

3) Sample gain variation

$$x_3 = \sigma_R / \overline{R} \tag{10}$$

where

$$\overline{R} = \langle R(n) \rangle \tag{11}$$

and

$$\sigma_R = \langle [R(n) - \overline{R}]^2 \rangle^{1/2}. \tag{12}$$

4) Spectrally related features (reflection coefficient averages)

$$x_{i+3} = \langle k_i(n) \rangle \quad \text{for } i = 1, 2, \cdots, M. \tag{13}$$

The *feature vector* x is defined by

$$x^T = [x_1 x_2 \cdots x_{3+M}] \tag{14}$$

where T denotes transpose.

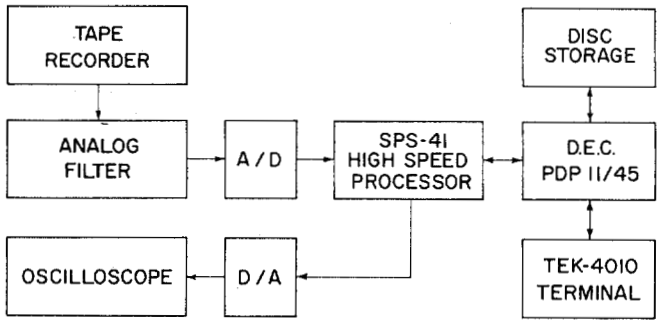

Fig. 1. Block diagram of system for processing speech.

III. Procedures

A. Data

The data used for the analysis were obtained during interviews of four speakers. Each interview was then edited so that only the interviewees' voices remained. The total duration of each edited interview (including pauses) was typically 15-18 min. The total data base used for this study was approximately one hour in duration. No special precautions or recording conditions were imposed on the experiment. Interviews were conducted in normal room environments with a dynamic microphone and an audio tape recorder. So that a small number of speakers could be used with some generality in extrapolating results, a homogeneous population of four male speakers was chosen, each having somewhat similar speech characteristics and relatively narrow fundamental frequency ranges. Histograms of the raw nonaveraged fundamental frequency values showed substantial overlap among the four speakers.

B. Digital Processing of Data

The audio tape was digitally processed using the system shown in Fig. 1. Each test segment was recorded onto a disk using conventional procedures. A novel part of the procedure is based upon the use of a high-speed signal processor and oscilloscope (for visual feedback during processing). Using an array-processing software system, it is possible to process the data in real time at a 50 Hz analysis frame rate from a Fortran environment. Processing includes modified cepstral pitch period and voicing detection, gain calculation, linear prediction analysis for reflection coefficients, and a running mean and mean-square computation of these parameters.

The procedure for generating output feature vectors to be used in the statistical analysis is shown in Fig. 2. A counter for frame n is incremented and one frame of speech is analyzed. The parameters used are: sampling frequency f_s = 6.5 kHz, number of analysis coefficients $M = 10$, number of samples for reflection coefficient computation = 128, and the number of samples for F_0, and gain parameter analysis = 256 (40 ms). The analysis frame rate is 50 Hz. Preemphasis of the speech data is applied using a differencer, $1 - z^{-1}$.

Fundamental frequency estimation is performed with a modified cepstral technique. After the spectrum has been computed, a symmetrical window function is applied that smoothly tapers from unity at 1000 Hz to zero from 1500 Hz to $f_s/2$. This simple modification resolves most of the voicing problems one obtains with the usual cepstral analysis method since only

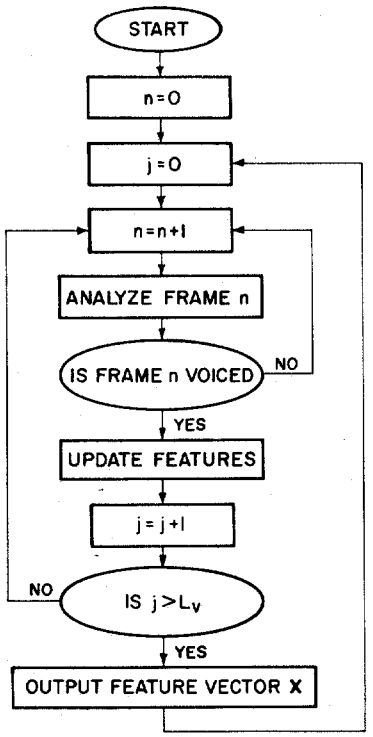

Fig. 2. Procedure for generating output feature vectors.

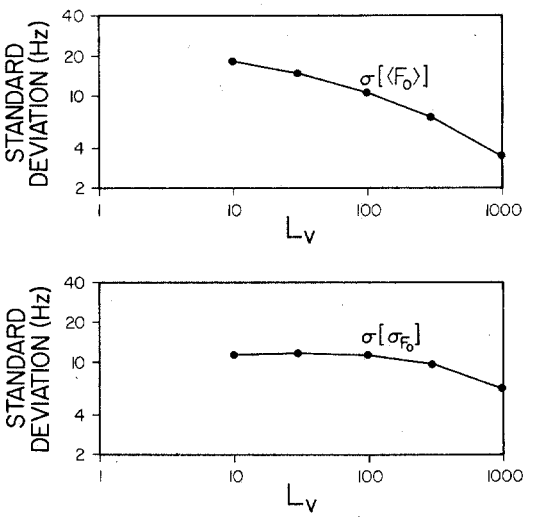

Fig. 3. Standard deviation of F_0 related features as a function of the number of voiced frames L_v.

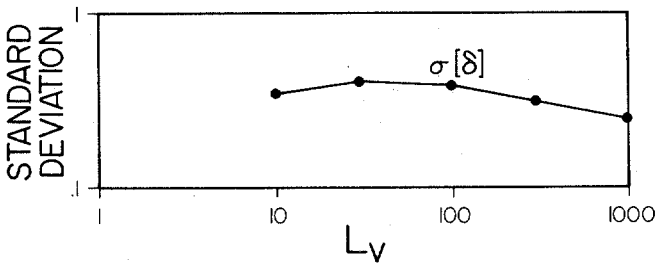

Fig. 4. Standard deviation of gain related features as a function of the number of voiced frames L_v.

the most consistent region of harmonic structure is used [3]. Two frames of delay are included in the system so that some amount of error detection and correction can be applied in the pitch period estimation. One additional test has been found necessary for obtaining meaningful feature vectors. A max (F_0) and min (F_0) value are chosen for the speaker being analyzed to ensure against gross errors causing the fundamental frequency features from being dramatically affected. If $\min(F_0) < F_0 < \max(F_0)$, the frame is judged to be voiced and the long-term averages are updated. The frame counter is incremented and if $l > L_v$, the resultant features vector x is output to disk, l is reset to zero, and analysis then continues.

IV. Experiments

A. Experiment 1—Statistical Variation as a Function of L_v

The complete edited audio tape for speaker D (approximately 18 min in duration) was analyzed to extract long-term averaged feature vectors for several L_v conditions. As a time reference, $L_v = 1000$ corresponds to approximately 70 s. The total number of vector samples obtained is approximately inversely proportional to L_v.

The unbiased variance estimate of the feature $x = \langle p(i) \rangle$ based upon the speech parameter p is

$$\sigma^2(x) = \frac{1}{L_f - 1} \sum_{i=0}^{L_f - 1} [\langle p(i) \rangle - \bar{x}]^2 \tag{15}$$

where

$$\bar{x} = \frac{1}{L_f} \sum_{i=0}^{L_f - 1} \langle p(i) \rangle. \tag{16}$$

Each $p(i)$ explicitly denotes an individual feature, and L_f is the number of feature vectors obtained over the total speech duration. Note that L_f is actually a function of L_v since the total duration is fixed. The sample mean x is thus independent of L_v except for sampling variation in the real-time analysis because it is not possible to start analysis at precisely the same location on the audio tape when L_v is changed. The true variance σ_p^2 is estimated from $\sigma_p^2 = \sigma^2(x)$ with $L_v = 1$. Features which themselves are based on variances (such as $x_2 = \sigma_{F_0}$ and $x_3 = \sigma_R/\bar{R}$) do not allow for a true variance estimate. The sample standard deviations of the fundamental frequency-related features are shown in Fig. 3 as a function of L_v. The estimated standard deviation about the long-term fundamental frequency averages is reduced from about 18 Hz for $L_v = 10$ to about 6 Hz for $L_v = 1000$. These values are somewhat higher than the long-term F_0 averages reported by Horii [8]. However, this experiment is based upon unconstrained conversational speech, whereas Horii's experiment was based upon a reading of the "Rainbow Passage."

The estimated standard deviation of the $x_2 = \sigma_{F_0}$ feature is surprisingly constant, until at least on the order of 7-10 s of speech ($L_v > 100$) have been analyzed. Increasing L_v from 100 to 1000 decreases $\sigma(x_2)$ from 12-6 Hz.

The variability of the gain variation feature, $\sigma(x_3)$, as shown in Fig. 4, follows a similar pattern. This particular feature appears to be a very weak function of L_v.

In Fig. 5(a), the estimated standard deviation of the $\langle k_1 \rangle$ and $\langle k_{10} \rangle$ features for speaker D are shown as being representative of the reflection coefficient feature set characteristics. Although the estimated standard deviation does not decrease as

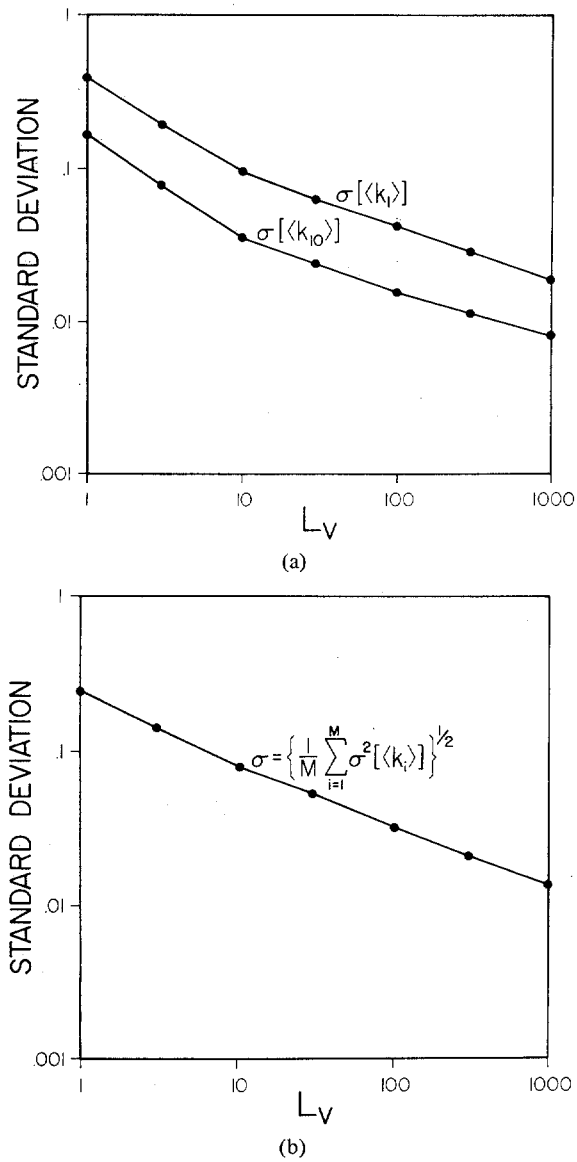

Fig. 5. (a) Standard deviation of reflection coefficient averages as a function of the number of voiced frames L_v. $\langle k_1 \rangle$, $\langle k_{10} \rangle$ deviations. (b) Standard deviation of reflection coefficient averages as a function of the number of voiced frames L_v. rms of all coefficient variances.

rapidly as predicted by sampling theory for the case of independent samples because of intraspeaker variability, the decrease is substantial and is surprisingly linear on a log-log scale. Instead of a $L_v^{-1/2}$ relation, the standard deviation of the reflection coefficient features appears to approximately decrease proportionally to a $L_v^{-1/3}$ model beyond $L_v = 10$.

The rms deviation over all $\langle k_i \rangle$ averages is shown in Fig. 5(b). Over a range of L_v from 10 to 1000, the $L_v^{-1/3}$ model is still seen to be very accurate for predicting the decrease in reflection coefficient feature parameter variation as L_v is increased. The measured exponent value is certainly dependent upon the particular speaker. However, it appears to vary only slightly from the model discussed for the several other speaker measurements.

The estimate of the true variance for the k_1, k_{10}, and overall parameter variance is also shown in Fig. 5(a) and (b) at $L_v = 1$.

A second way of qualitatively showing the effect of long-term averaging is to show two-dimensional scatter diagrams for

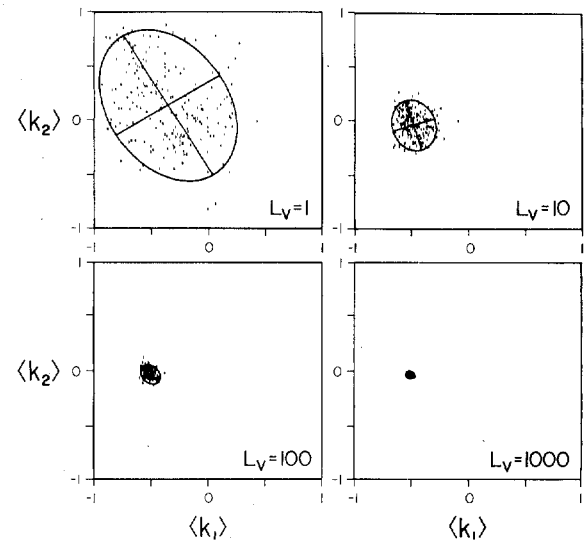

Fig. 6. Scatter plots of $\langle k_1 \rangle$, $\langle k_2 \rangle$ features for different L_v with two-sigma ellipses and principal axes.

various values of L_v. Fig. 6 shows a scatter plot of $\langle k_2 \rangle$ versus $\langle k_1 \rangle$ samples. Each point is based upon L_v samples from the edited audio tape for speaker D. Shown with the data are two-sigma ellipses with the principal axes. A dramatic decrease in the dispersion of the data is seen as L_v increases.

B. Experiment 2—Discrimination as a Function of L_v

The approach taken here is to investigate the effectiveness of long-term averaging for speaker recognition using the ratio of the between-speaker variance and the within-speaker variance, without specifying particular speaker recognition experiments. Since the mathematics of this procedure (Fisher discriminant method) is discussed elsewhere [1], only the necessary details will be summarized below.

A within-speaker covariance matrix W is computed, and then a normalized between-speaker covariance matrix B' is found in terms of the matrix B of Bricker *et al.* [1] from

$$B' = B/L_f \tag{17}$$

where L_f is the number of feature vectors. The normalization is included so that B' will depend only upon the sample means, not upon the number of feature vectors. Eigenvalues and eigenvectors of the equation

$$B'\boldsymbol{\phi}_k = \lambda_k W \boldsymbol{\phi}_k \tag{18}$$

are then obtained. The eigenvalues are ordered from highest to lowest, and as the number of speakers, four, is less than the number of features, thirteen, all but the first three eigenvalues are zero [1]:

$$\lambda_1 \geqslant \lambda_2 \geqslant \lambda_3 \geqslant \lambda_4 = \lambda_5 = \cdots = \lambda_{13} = 0. \tag{19}$$

A new coordinate system is defined using the eigenvectors of (18) as base vectors, so that the new coordinate vector y is related to x through the linear transformation

$$y = \boldsymbol{\phi}^T x \tag{20}$$

where $\boldsymbol{\phi}$ is the matrix whose columns are the eigenvectors of (18). The eigenvalues of (18) represent the variance ratios in the directions of the eigenvectors with λ_1 being the maximum

TABLE I
VARIANCE RATIOS OF LONG-TERM AVERAGE FEATURE SET FOR $L_v = 100$ AND $L_v = 1000$

FEATURES	VARIANCE RATIO	
	$L_V = 100$	$L_V = 1000$
$x_1 = \langle \bar{F}_0 \rangle$	0.332	2.321
$x_2 = \langle \sigma_{F_0} \rangle$	0.004	0.043
$x_3 = \langle \delta \rangle$	0.119	0.329
$x_4 = \langle k_1 \rangle$	0.081	0.305
$x_5 = \langle k_2 \rangle$	2.721	16.118
$x_6 = \langle k_3 \rangle$	0.221	1.216
$x_7 = \langle k_4 \rangle$	0.367	2.023
$x_8 = \langle k_5 \rangle$	0.307	2.002
$x_9 = \langle k_6 \rangle$	2.315	11.452
$x_{10} = \langle k_7 \rangle$	0.155	0.650
$x_{11} = \langle k_8 \rangle$	0.511	2.591
$x_{12} = \langle k_9 \rangle$	0.185	0.977
$x_{13} = \langle k_{10} \rangle$	0.403	0.978

TABLE II
VARIANCE RATIOS OF TRANSFORMED LONG-TERM AVERAGE FEATURE SET FOR $L_v = 100$ AND $L_v = 1000$

TRANSFORMED FEATURES	VARIANCE RATIO	
	$L_V = 100$	$L_V = 1000$
y_1	10.959	115.368
y_2	0.972	7.956
y_3	0.393	1.730
y_4	0	0
⋮	⋮	⋮
y_{13}	0	0

variance ratio, λ_2 being the next largest (in a direction orthogonal to ϕ_1), etc. Variance ratios can also be computed in the original coordinate system as a method for measuring relative effectiveness of features.

Tables I and II show the variance ratios in the original and transformed coordinate systems, respectively, for $L_v = 100$ and $L_v = 1000$. Except for the fact that parameter correlation is not taken into account, the variance ratio values can be taken as quantitative measures of the original parameter's effectiveness in speech recognition. For example, we see that σ_{F_0} provides very little discrimination among speakers, whereas $\langle k_2 \rangle$ appears to provide the maximum discrimination among speakers over all parameters. It is seen that the first dimension in the new coordinate system results in a substantially increased variance ratio.

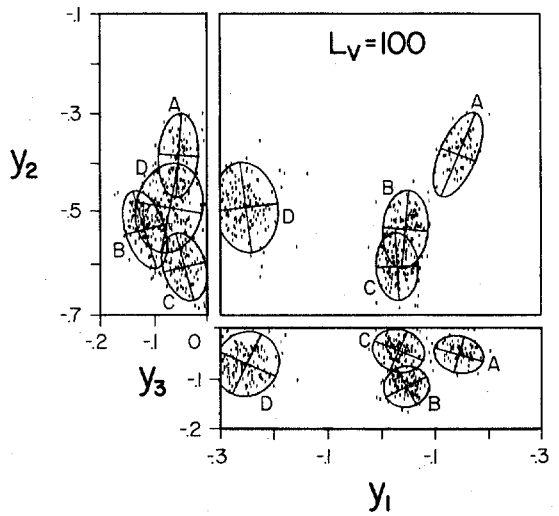

Fig. 7. Scatter plots for speakers *A*–*D* along first three Fisher discriminant dimensions ($L_v = 100$).

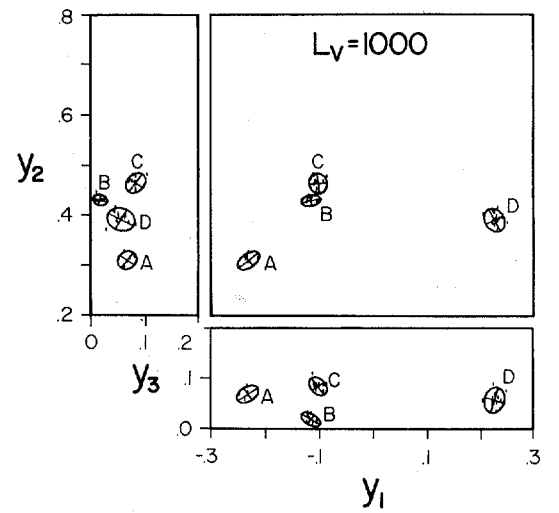

Fig. 8. Scatter plots for speakers *A*–*D* along first three Fisher discriminant dimensions ($L_v = 1000$).

Two-dimensional scatter plots of the first three transformed dimensions are shown in Fig. 7 for $L_v = 100$ and in Fig. 8 for $L_v = 1000$. The results are based upon the four speakers *A*–*D*. Also shown are two-sigma ellipses and the principal axes for each speaker distribution. In Fig. 7 it is seen that *A* and *D* are essentially uniquely separated from *B* and *C* in at least one plane ($y_1 - y_2$). A relatively large overlap does occur, however, for *B* and *C* in all planes. A cursory comparison of Fig. 7 and the relative sizes of clusters in Fig. 6 will illustrate that substantial benefits in discriminating against different speakers have been obtained over using no averaging ($L_v = 1$) or very limited amounts of averaging ($L_v = 10$).

In Fig. 8, it is seen that by performing long-term averaging with $L_v = 1000$, perfect discrimination is obtained, in this instance based upon only a two-dimensional transformed feature representation.

The variance ratios for the input feature variables are shown in Table I for $L_v = 100$ and $L_v = 1000$. If the variables were statistically independent, these ratios would differ by a multi-

plicative factor of 10 rather than the smaller factors indicated in the figure. The ordering of the features in terms of variance ratios is of some interest. The fifth feature, k_2, clearly shows the largest variance ratio, with the ninth feature, k_6, the next largest. These coefficients correspond to the coefficients for the highest power of z^{-1} in the models of order 2 and 6 found from linear prediction analysis. The two-pole model has been used in earlier recognition tasks [18].

The variance ratios for the first three features, fundamental frequency, its standard deviation, and sample gain variation, are smaller than what one might expect from intuition. Part of the reason may lie in the fact that the speakers were chosen to have similar fundamental frequency ranges.

The variance ratios for the new coordinate system, the eigenvalues of (18), are shown in Table II. From these ratios and the scatter diagrams of Fig. 8, it can be seen that very clear separation of the speakers is indicated for the long-term average case of $L_v = 1000$ by using only the first two coordinates, y_1 and y_2, in the direction of the eigenvectors $\boldsymbol{\phi}_1$ and $\boldsymbol{\phi}_2$.

V. Discussion

A. Parameter Variability Over Days, Weeks, Etc.

This initial study has been restricted to the study of long-term averages taken from one session. This is probably the reason why the standard deviation of the long-term averages tends to have a monotonically decreasing behavior. Although some amount of intraspeaker variability is reflected in the data, additional variability will occur when results are obtained from sessions separated by days, weeks, or months later. In several studies over linguistically constrained units, this effect has been shown to be severe beyond several months for short text-dependent segments [4]. A large data base extending over several months is now being generated for studying these effects in conversational speech.

B. Accuracy of Voicing Decisions

Since all long-term statistics are made only during voicing, it is very important to know that realistic voicing decisions are made. Spectral slope and normalized gain variation are direct computations requiring no decisions (except for voicing) and are, therefore, very robust.

If the threshold setting for voicing and pitch period detection is set too high or too low, the effect can be catastrophic. At one extreme, if the voicing threshold is too high, very few frames will be included in the statistics as being voiced (although they will be very reliable estimates) and, furthermore, transitions in which considerable fundamental frequency variations may occur are likely to be missed, causing the measured fundamental frequency standard deviation to be unrealistically small.

At the other extreme, if the threshold is too low, there will be a tendency to define fundamental frequencies near the maximum allowable frequency (minimum pitch period) (near 400 Hz) during actual voiced speech and at random values throughout the rest of the allowable range during unvoiced speech. Although a pitch period and voicing decision program with several frames of delay is used for error detection and correction, it is essentially impossible to separate accurate estimates from gross errors beyond some reasonable threshold.

C. Assumptions Versus Experimental Results

It was assumed that $\langle F_0 \rangle$ carries important speaker information. The $\sigma[\langle F_0 \rangle]$ versus L_v graph in Fig. 3 showed a significant monotonic decrease as L_v was increased. In addition, the variance ratio was relatively high (even though speakers were purposely chosen with similar fundamental frequency ranges). Therefore, this assumption appears valid. The assumption that $\sigma[\sigma_{F_0}]$ is meaningful does not appear to be true for conversational speech. The variance ratio for this feature is extremely small. This result contradicts that shown by Mead [13], where the use of the first through the fourth moments of F_0 and of the first four differences of F_0 (resulting in 20 features) was suggested. Our experience indicates that unless hand-marked or hand-corrected F_0 contours are used, very significant biases in results can occur because of very few gross errors in F_0 estimation. Higher order differences and moments only magnify these biases.

The standard deviation of the gain deviation feature as a function of L_v shows a weak relationship to expectations from statistical sampling theory. In addition, the variance ratios for the gain deviation feature are relatively small. Although some discrimination is obtained, what we have seen is that not only is there substantial intraspeaker variability for this parameter, but that, in addition, considerable overlap in the gain feature values occurs between speakers. Other measures of fundamental frequency and gain variations may prove to be more useful than the ones used here, which are essentially based upon root mean squares taken about the averages. One possibility is the use of the ratio of geometric and arithmetic means as used in evaluating spectral flatness [11].

The long-term averages of the reflection coefficients as a set appear to be the most significant features for speaker recognition. Not only does the standard deviation of the long-term averages show a substantial decrease as a function of L_v, but in addition, the variance ratios are seen to be relatively large for most of the parameters.

D. Observations on Reflection Coefficient Averaging

Although $\sigma(\langle k_{10} \rangle) < \sigma(\langle k_1 \rangle)$ for all L_v, in Fig. 5(a), one should not be misled into thinking that $\langle k_{10} \rangle$ is a better feature for speaker recognition. This result occurs because k_1 inherently has a larger standard deviation than k_{10} ($\langle k_1 \rangle = k_1$ for $L_v = 1$). The important fact to note is that whatever the parameter deviation is without averaging, due to either linguistic content or intraspeaker variability, it decreases as $L_v^{-\alpha}$ where $\frac{1}{3} \leqslant \alpha \leqslant \frac{1}{2}$ when long-term averaging is applied.

In a recent paper [17], the use of orthogonal linear prediction parameters for use in text-independent speaker recognition studies was suggested. Although very high recognition scores were shown using the orthogonal linear prediction parameters, we would suggest that substantial reduction in scores would occur if unconstrained data bases as described here were used. Whatever scores are obtained using, in effect, $L_v = 1$,

our results qualitatively indicate that substantial improvements could occur by incorporating long-term averaging.

Each orthogonal parameter was obtained from a linear combination of all reflection coefficients as

$$\Phi_i = \sum_{j=1}^{M} c_{ij} k_j \tag{21}$$

where the c_{ij} terms were obtained from a principal component analysis. The averaged parameters would then be

$$\langle \Phi_i \rangle = \sum_{j=1}^{M} c_{ij} \langle k_i \rangle. \tag{22}$$

Although Fig. 6 shows only the dispersion characteristics for $\langle k_1 \rangle$ and $\langle k_2 \rangle$, similar characteristics are obtained for all the coefficients. The amount of data dispersion will be primarily due to the value of L_v, not the fact that a linear combination of the k_i terms (or the $\langle k_i \rangle$ terms) has been obtained.

E. Computational Considerations

Studies of this type place a premium on the available processing speed of the computer system. It became clear early in the study that small- or medium-scale computer capability was insufficient. For example, the analysis method described runs in approximately 100 times real time if all operations are implemented only on the PDP-11 system. The relatively small data base of speakers for this study would have required over 100 hours of processing time.

Except for the nontrivial costs in software development, we have found that attaching a high-speed processor to the main computer system provides a very economical solution to the requirements for real-time processing.

VI. Summary

The properties of long-term feature averaging for three sets of fundamental frequency related, gain related, and spectrally related parameters have been investigated. Based upon the Fisher discriminant method, the rank ordering of the parameter sets in importance was shown to be spectral, fundamental frequency, and then gain. It was also shown that over a long duration from $L_v = 10$ to $L_v = 1000$, the standard deviation of the sample means of the reflection coefficient vectors decreased proportionally to $L_v^{-1/3}$.

A small number of speakers with relatively homogeneous characteristics was used to illustrate the effects of long-term averaging. The data base was of nontrivial duration, somewhat greater than one hour in length. Furthermore, the text was unconstrained conversational speech, recorded under normal room noise conditions. Analysis was performed in real time with a high-speed signal processor.

Presently, other spectral representation methods are being investigated and a data base is being developed for performing text-independent speaker recognition tests without any linguistic or structural constraints.

References

[1] 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.

[2] G. R. Doddington, "A method of speaker verification," Ph.D. dissertation, Univ. of Wisconsin, Madison, 1970.

[3] O. Fujimura, "An approximation to voice aperiodicity," *IEEE Trans. Audio Electroacoust.*, vol. AU-16, pp. 68–72, 1968.

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

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

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

[7] J. N. Holmes, "Low-frequency phase distortion of speech recording," *J. Acoust. Soc. Amer.*, vol. 58, pp. 747–749, 1975.

[8] Y. Horii, "Some statistical characteristics of voice fundamental frequency," *J. Speech Hearing Res.*, vol. 18, pp. 192–201, 1975.

[9] U. Kosiel, "Statistical analysis of speaker-dependent differences in the long-term average spectrum of Polish speech," in *Speech Analysis and Perception*, vol. 3, W. Jassem, Ed., Polish Academy of Sciences. Warsaw, Poland: PWN-Polish Scientific Publishers, 1973, pp. 117–120.

[10] R. C. Lummis, "Speaker verification by computer using speech intensity for temporal registration," *IEEE Trans. Audio Electroacoust.*, vol. AU-21, pp. 80–89, 1973.

[11] J. D. Markel and A. H. Gray, Jr., *Linear Prediction of Speech*. Berlin, Heidelberg, New York: Springer-Verlag, 1976.

[12] J. D. Markel and D. Wong, "Considerations in the estimation of the glottal volume velocity waveforms," submitted to *IEEE Trans. Acoust., Speech, Signal Processing.*

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

[14] W. S. Mohns, "Statistical feature evaluation in speaker identification," Ph.D. dissertation, North Carolina State Univ., Raleigh, 1969.

[15] A. E. Rosenberg, "Effect of glottal pulse shape on the quality of natural vowels," *J. Acoust. Soc. Amer.*, vol. 49, pp. 583–590, 1970.

[16] M. Rothenberg, "A new inverse-filter technique for deriving the glottal air flow waveform during voicing," *J. Acoust. Soc. Amer.*, vol. 53, pp. 1632–1645, 1973.

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

[18] M. R. Sambur and L. R. Rabiner, "A speaker independent digit recognition system," *Bell Syst. Tech. J.*, vol. 54, pp. 81–102, 1975.

[19] K. N. Stevens, "Sources of inter- and intra-speaker variability in acoustic properties of speech sounds," in *Proc. 7th Int. Congr. of Phonetic Sciences*, A. Rigault and R. Charbonneau, Ed. The Hague: Mouton, 1972.

[20] K. N. Stevens, C. E. Williams, J. R. Carbonell, and B. Woods, "Speaker authentication and identification: A comparison of spectrographic and auditory presentation of speech material," *J. Acoust. Soc. Amer.*, vol. 44, pp. 1596–1607, 1968.

[21] W. D. Voiers, "Perceptual bases of speaker identity," *J. Acoust. Soc. Amer.*, vol. 36, pp. 1965–1973, 1964.

[22] J. J. Wolf, "Efficient acoustic parameters for speaker recognition," *J. Acoust. Soc. Amer.*, vol. 51, pp. 2044–2056, 1972.