Automatic Speaker Recognition Based on Pitch Contours*†

B. S. Atal

Bell Telephone Laboratories, Incorporated, Murray Hill, New Jersey 07974

(Received 31 March 1972; revised 12 July 1972)

An automatic speaker-recognition method, using temporal variations of pitch in speech as a speaker-identifying characteristic, is described. The pitch data was obtained from 60 utterances, consisting of six repetitions of the same sentence, spoken by 10 speakers. The pitch data for each utterance was represented by a 20-dimensional vector in the Karhunen–Loève coordinate system. The 20-dimensional vectors representing the pitch contours were linearly transformed so that the ratio of interspeaker to intraspeaker variance in the transformed space was maximized. A reference vector was formed for each speaker by averaging the transformed vectors of that speaker. The recognition procedure was based on measuring the Euclidean distance between the test vector and the reference vectors in the transformed space; the speaker corresponding to the reference vector with the smallest distance was selected as the speaker of the test utterance. The percentage of correct identifications was found to be 97%. The results suggest that temporal variations of pitch could be used effectively for automatic speaker recognition.

Subject Classification: 9.10, 9.3.

INTRODUCTION

The ability of human listeners to identify speakers from their voices has long been known. Although the estimates of how reliably humans perform this task vary,[1–3] it is generally agreed that they do it very effectively. It is therefore no wonder that there has been a great interest recently in evolving automatic (computer) methods of voice identification, which will match or even surpass human performance. Automatic methods of identifying or verifying speakers from their voices may have potential applications in many diverse fields. Speaker verification techniques, for example, could be employed by the banking and business world to provide new and almost revolutionary services to their customers. These techniques could be used to provide controlled access to a facility or information to selected individuals. Reliable speaker identification by voice can be extremely useful when other clues to the speaker's identity are either missing or highly ambiguous.

Several recent studies [4–10] have demonstrated that automatic speaker recognition, at least within a small population, is indeed feasible. The present study differs from these previous studies in two important respects. First, in most of the past studies, the speech signal was transformed into the spectral form and the resulting time–frequency-energy (spectrographic) patterns were used to identify or verify the speaker. There is no doubt that the spectral characteristics of speech are important for speaker recognition. However, there are other speech characteristics, such as pitch, which are missing from the spectra but which could be useful in distinguishing one speaker from another. Pitch information has important advantages over spectral information for speaker recognition. For example, spectral patterns are affected by the frequency characteristics of the transmission system. Thus, some kind of normalization is required to eliminate the influence of any variable transmission characteristics on the spectral data. Such normalization is often difficult to achieve. Spectral data also depends upon the level at which the speaker talks and the distance between the talker and the microphone. Pitch, on the other hand, is unaffected by such variations. Second, in this study, the entire pitch contour of a sentence-length utterance is used for speaker recognition—not just the average pitch of the speaker. Such time-dependent information has important advantages. For example, an impostor may be able to mimic those voice characteristics of a speaker which remain fixed in time, such as the average pitch or the time-averaged spectrum. However, it appears improbable that an impostor could easily mimic the entire variation of pitch as a function of time.

On a more fundamental level, the voices of two persons differ due to the physical differences between their vocal organs and the manner in which they use them during speech production. The parts that principally determine the spectra are the vocal cavities and the articulators; the characteristics of the vocal cords are reflected in the pitch information. The vocal cords play an important role in the production of voiced sounds. During speech production we continually alter the ten-

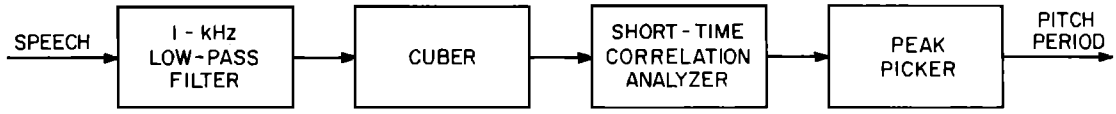

FIG. 1. Block diagram of the pitch detector.

sion of the vocal cords and the subglottal air pressure to achieve the desired pitch variation with time. Speaker recognition on the basis of pitch is feasible if the pitch patterns of different speakers are distinct from each other. It seems quite likely that this is the case.

The aim of the present study was to find the importance of pitch for automatic speaker recognition. The task of speaker recognition could be either that of identifying a person from a population of several known speakers or that of verifying whether a person is what he claims to be. Since the objective of this study was to determine if the pitch contours of individuals were different, the speaker identification task, being more sensitive in showing such differences, was selected for conducting speaker-recognition tests. The complete investigation was broadly divided into two parts. The first part of this paper describes the data-collection procedure, which included the speech recording, the analog-to-digital conversion of the speech samples, and the pitch analysis. The pitch analysis is discussed only briefly in this paper; it is described in detail in Refs. 11 and 12. The second part describes the recognition procedure, the data-reduction methods, and the results of the speaker-recognition tests.

I. SPEECH RECORDING

Speech recordings were made in an anechoic chamber, using a Brüel & Kjær 0.5-in. condenser microphone. Ten female speakers were selected to read the specified text. It was considered desirable that the speakers be of the same sex. Persons with widely different average pitch (as is likely to be the case with men and women) can be easily distinguished from one another on the basis of their average pitch alone and consequently no advantage is gained by selecting speakers of both sexes. The average duration of pitch periods for the 10 speakers ranged between 4.3 and 5.6 msec with a mean of 4.8 msec and a standard deviation of 0.4 msec. The 10 speakers will be referred to by their initials BB, NC, IE, PL, JO, JS, LB, CW, MS, and JH, respectively, in the paper.

Each speaker read a set of five sentences written on a card. Only one of the five sentences, namely, "May we all learn a yellow lion roar," having the advantage of being entirely voiced, was used in the speaker-recognition tests. The duration of this sentence varied between 1.8 and 2.8 sec among the speakers. The other sentences were read by the speakers in order to avoid any special emphasis on the key sentence. The order in which the sentences were read was different for each of the six recordings for every speaker. The speakers were not given any instructions about the manner in which they should read the sentences. They were, however, told that the recordings would be used in future speaker-recognition tests. Recordings were made on two different days with an interval of 27 days apart; on each day, the recordings were made in three separate sessions during a 30-min period providing a total of 60 utterances for the 10 speakers.

II. PITCH ANALYSIS

Prior to pitch analysis, the analog speech signal was converted into digital form for processing on a digital computer. In the analog-to-digital converter, the speech signal was filtered by a low-pass filter having an attenuation of 3 dB at 4 kHz and more than 40 dB above 5 kHz, sampled sequentially at a rate of 10 000 times/sec, and quantized into 12-bit binary numbers. In the computer, the durations of the individual pitch periods in each of the 60 utterances were determined. An automatic method for performing the pitch period analysis was developed.[11,12] In this method, the speech signal was filtered through a 1-kHz low-pass filter and each sample of the low-pass filtered signal was raised to the third power to emphasize the high-amplitude portions of the speech waveform. The duration of the pitch period was obtained by performing a short-time correlation analysis on the cubed low-pass filtered speech. A block diagram illustrating the pitch analysis is shown in Fig. 1.

Two examples of the pitch contours obtained by the above method for each of the 10 speakers are shown in Fig. 2. [13] The speakers are identified by their initials. The ordinate is the duration of the pitch period in milliseconds and the abscissa represents the time in seconds. In most cases, the pitch contours of the same speaker were found to be similar. Comparatively large jumps in the durations of the pitch periods frequently occurred at the onset and end of voicing. However, they were not found to recur consistently among the different utterances of the same speaker. The same could also be said for the period-to-period variations of pitch.

The durations of the pitch contours were found to vary from one utterance to another, even for the same speaker. It was considered desirable to normalize the duration of each utterance to the same time interval and, if necessary, to use duration as an additional measure for speaker recognition. A new set of time coordinates was computed for each utterance by linear scaling of the original time coordinates such that the

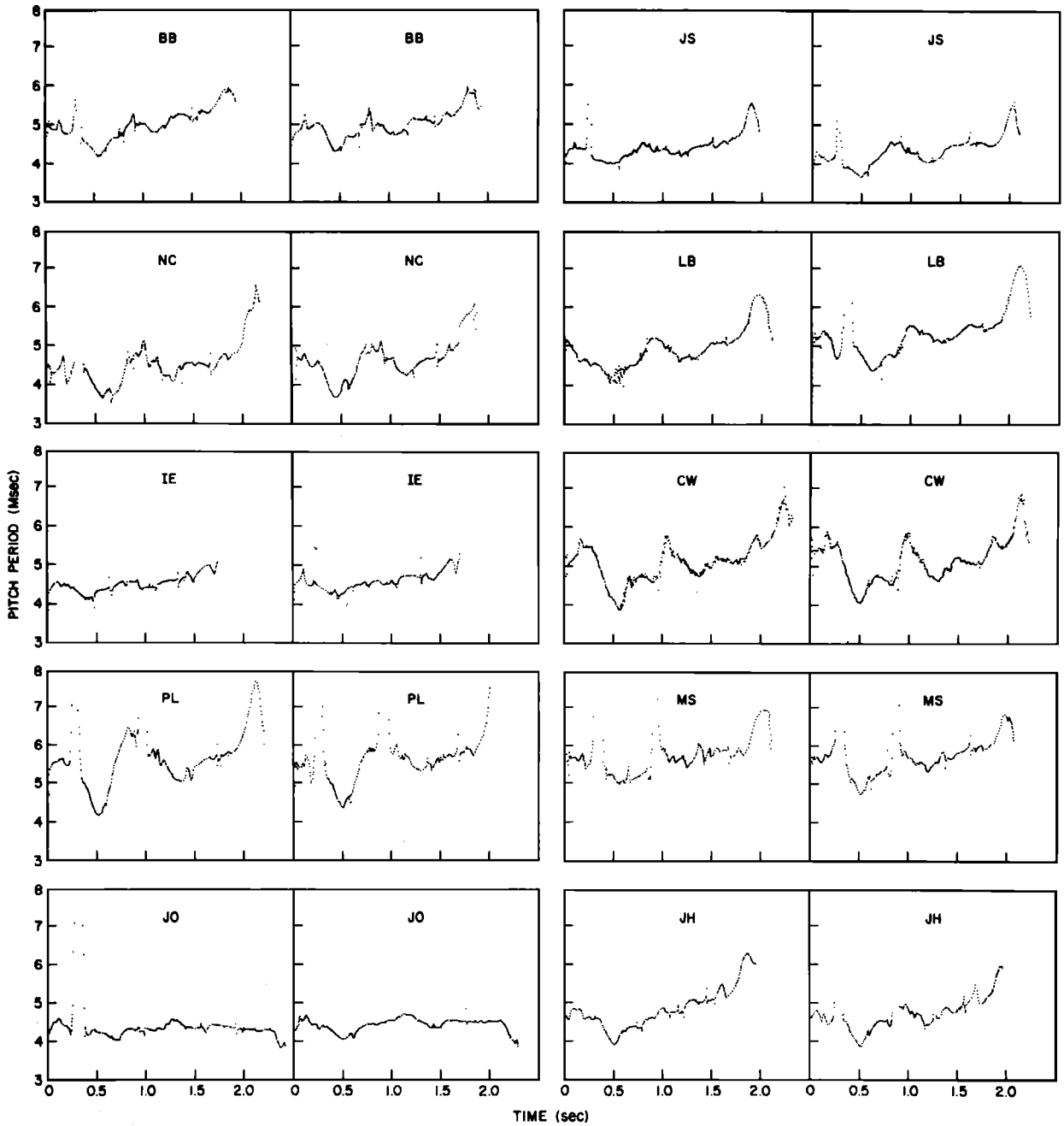

FIG. 2. Some examples of the pitch contours.

total duration of the utterance was 2 sec. For the utterances having no interruptions in voicing, the duration was assumed to be the difference between the time coordinates of the last and first pitch periods; otherwise, the duration was assumed to be equal to the sum of the durations of the individual voiced segments. There were at most two interruptions in any utterance; in every case they occurred either after the word "we," or after the word "learn" and were easily identified.

III. SPEAKER RECOGNITION PROCEDURE

The problem of developing efficient classification procedures can be approached from two different standpoints. One can attempt to find, by visual inspection of the data, a set of distinctive features that distinguish the pitch data of one speaker from those of another. These features can then be used as the basis for classification of an unknown speaker. Alternatively, the abstraction of distinctive features and the development

of decision rules can be done on a computer by means of algorithms based on statistical decision theory. Although both methods have merit, the second is better suited for processing large amounts of data. Consequently, in our work, the tools of statistics were employed to the fullest extent wherever it was feasible to do so.

Statistical classification techniques (sometimes called statistical pattern recognition procedures) usually cover three kinds of situations: (1) The probability distributions of the measures characterizing the different classes are known; (2) only the forms of the distributions are known; and (3) nothing is known about the distributions. The speaker-classification problem falls into the third category. The probability distributions are neither known nor can they be determined from the available data in any meaningful sense. There are at least two ways in which the problem of unknown distributions is usually approached. Either a nonparametric classification technique is used, which does not require the knowledge of the distributions, or a specific form of the distribution is assumed and the decision rules are chosen to be optimum for that distribution. There is no clear advantage in using the second approach. In our particular case, it is not necessary to know the distributions in order to find an efficient classification procedure; only the boundaries separating the regions for each class in the measurement space need be known and these can be determined by direct search. The effectiveness of a classification procedure based on a particular boundary can be obtained from the error rates in the speaker-recognition tests. Another advantage of the direct search procedure is that the search can be terminated at the simplest boundary which works well enough for the given purpose. Simple boundaries, such as hyperplanes, have the advantage that they are usually easier to implement. Furthermore, the size of the input data may well restrict the choice of boundaries. For example, in the present problem of speaker classification, one may represent each utterance of a speaker by a point in an N-dimensional Euclidean space. It is readily seen that a total of N utterances of any two speakers can be exactly separated into two classes by a hyperplane in this space. The search for boundaries more complex than the hyperplane is thus unnecessary unless the number of utterances per speaker is significantly larger than $N/2$.

A. Formation of Design and Test Sets

The entire pitch data, consisting of six utterances per speaker, was partitioned into design and test sets. The utterances in the design set were used to find the optimum separating boundaries between the different speakers; the utterances in the test set were employed to test the effectiveness of the boundaries for speaker recognition. This procedure is necessary since the separability of the classes in the design set does not guarantee that the independent utterances of the test set are also separable. In fact, pattern recognition procedures, adjusted to work remarkably well on the design samples, might often perform very poorly on new samples of the test set. The reasons for the discrepancy are not hard to find. With only a finite amount of data, it is relatively easy to find a procedure which could classify all the patterns of the design set without any error. However, the procedure will be effective on the test data only if it employs the characteristics common to all members of any class and not just those represented in the design set.

In the present case, the number of samples available for the design and test sets was limited. To make full use of the available samples, each of the six utterances for every speaker in turn was used as a test sample, while the remaining five utterances were used to design the classification procedure. The above procedure maintains the independence of design and test sets, while permitting the use of all available utterances to test the effectiveness of the design.

B. Clustering Transformation

Consider the N-dimensional representation of the utterances (pitch contours), each utterance being represented by a point in an N-dimensional Euclidean space. The selection of the optimum boundary between any two classes is considerably simplified if points belonging to the same class (speaker) are clustered close to the centroid of the class. In this case, the concentration ellipsoids of various classes are approximately spherical in shape, and the proper dividing boundary between two classes is a hyperplane, which is the perpendicular bisector of the line joining the centroids of the concentration hyperspheres of the two classes. The task of finding a suitable "clustering" transformation plays a central role in the classification problem.[14]

The approach used in the present case is based on a generalization of the linear discriminant function, proposed by Fisher, to the multidimensional case.[15] In the latter case, we seek M best linear combinations of the vector components of each utterance which maximize the ratio of interclass to intraclass variance. The discriminant analysis also provides an estimate of the number of dimensions necessary to represent the interclass differences satisfactorily.

The clustering of the classes is produced by a linear transformation (matrix transformation) of the vector space. The linear transformation is determined as follows: First, a set of linearly independent vectors $\{\mathbf{u}_n\}$ is found which has the property that the linear combinations $\{z_{n,\alpha}^{(i)}\}$, defined by

$$z_{n,\alpha}^{(i)} \equiv \mathbf{u}_n{}^t \mathbf{y}_\alpha^{(i)}, \tag{1}$$

where $\mathbf{y}_\alpha^{(i)}$ is a vector representing the αth utterance of the ith speaker and $\mathbf{u}_n{}^t$ is the transpose of $\mathbf{u}_n$, will

separate the speakers maximally in the sense that the ratio of interspeaker to intraspeaker variance is maximized. Then, the linear transformation is given by the matrix U^t, where the nth column of the matrix U is the vector $\mathbf{u}_n$. It is easily shown that the vectors $\{\mathbf{u}_n\}$ are the solutions of the characteristic equation

$$B\mathbf{u}_n \equiv \lambda W\mathbf{u}_n, \tag{2}$$

where B and W are the interspeaker (between-speaker) and intraspeaker (within-speaker) covariance matrices, respectively.[7,16] $\mathbf{u}_n$ is an eigenvector and λ_n the corresponding eigenvalue of the characteristic equation. The matrices B and W are given by

$$B \equiv \langle [\bar{\mathbf{y}}^{(i)} - \bar{\mathbf{y}}^{(j)}][\bar{\mathbf{y}}^{(i)} - \bar{\mathbf{y}}^{(j)}]^t \rangle_{i,j}, \tag{3}$$

and

$$W \equiv \langle [\mathbf{y}_\alpha{}^{(i)} - \bar{\mathbf{y}}^{(i)}][\mathbf{y}_\alpha{}^{(i)} - \bar{\mathbf{y}}^{(i)}]^t \rangle_{\alpha,i} \tag{4}$$

where $\bar{\mathbf{y}}^{(i)} = \langle \mathbf{y}_\alpha{}^{(i)} \rangle_\alpha$ is the mean vector of the ith speaker, and $\langle\ \rangle$ indicates averaging over the indicated subscripts. Furthermore, the eigenvalue λ_n equals the ratio of interspeaker to intraspeaker variance for the linear combination represented by the eigenvector $\mathbf{u}_n$,[16] i.e., λ_n is a measure of the effectiveness of the nth linear combination for discriminating between different speakers.

Solution of the Characteristic Equation: Equation 2 is solved by converting it to a characteristic equation of a single real symmetric matrix.[17] Since W is a positive-definite symmetric matrix, it can be represented as

$$W = CLC^t, \tag{5}$$

where C is a unitary matrix ($C^{-1} = C^t$) whose columns are characteristic vectors of the matrix W,[18] and L is a diagonal matrix whose diagonal elements are the eigenvalues of W. On substituting Eq. 5 into Eq. 2, we obtain

$$B\mathbf{u}_n = \lambda_n CLC^t\mathbf{u}_n. \tag{6}$$

Since $CL^{-\frac{1}{2}}L^{\frac{1}{2}}C^t = I$ (identity matrix), Eq. 6 can be rewritten as

$$BCL^{-\frac{1}{2}}L^{\frac{1}{2}}C^t\mathbf{u}_n = \lambda_n CLC^t\mathbf{u}_n. \tag{7}$$

Next, we multiply both sides of Eq. 7 by $L^{-\frac{1}{2}}C^t$ and rewrite it as

$$L^{-\frac{1}{2}}C^tBCL^{-\frac{1}{2}}L^{\frac{1}{2}}C^t\mathbf{u}_n = \lambda_n L^{\frac{1}{2}}C^t\mathbf{u}_n. \tag{8}$$

Let us define a symmetric matrix S by

$$S = L^{-\frac{1}{2}}C^tBCL^{-\frac{1}{2}}, \tag{9}$$

and a vector $\mathbf{v}_n$ by

$$\mathbf{v}_n = L^{\frac{1}{2}}C^t\mathbf{u}_n. \tag{10}$$

Equation 8 can now be written as

$$S\mathbf{v}_n = \lambda_n\mathbf{v}_n. \tag{11}$$

The eigenvalues of S are identical to the eigenvalues of Eq. 2. The eigenvectors $\mathbf{u}_n$ are given from Eq. 10 by

$$\mathbf{u}_n = CL^{-\frac{1}{2}}\mathbf{v}_n. \tag{12}$$

The matrix U is then given by

$$U = CL^{-\frac{1}{2}}V, \tag{13}$$

where V is the modal matrix of S.

C. Recognition Procedure

The boundary separating any two speakers in the transformed space was chosen to be the hyperplane which was perpendicular to and passed through the center of the line joining the mean vectors of the two speakers. To classify an unknown vector $\mathbf{z}$ in the transformed space, its distance from the mean vector of each speaker was computed. The unknown vector was then assigned to the speaker whose mean vector was closest to the unknown. The distance between $\mathbf{z}$ and the mean vector of the ith speaker is given by

$$d_i = \|\mathbf{z} - \bar{\mathbf{z}}^{(i)}\|, \tag{14}$$

where $\|\ \|$ signifies the norm of the vector and is defined as the square root of the sum of the squares of the individual vector components. The decision rule assigns $\mathbf{z}$ to speaker j if $d_j < d_i$ for every i not equal to j. That the above decision rule is the exact equivalent of the hyperplane boundary can be seen from the following: Consider the separating boundary between any two speakers j and k. It is given by

$$\|\mathbf{z} - \bar{\mathbf{z}}^{(j)}\|^2 = \|\mathbf{z} - \bar{\mathbf{z}}^{(k)}\|^2. \tag{15}$$

Equation 15 can be rewirtten as

$$\mathbf{z}^t(\bar{\mathbf{z}}^{(k)} - \bar{\mathbf{z}}^{(j)}) = \tfrac{1}{2}[\|\bar{\mathbf{z}}^{(k)}\|^2 - \|\bar{\mathbf{z}}^{(j)}\|^2], \tag{16}$$

which is the equation of the desired hyperplane.

D. Data Reduction

It was necessary, for practical computational reasons, to reduce the number of samples by which each pitch contour was represented in the pitch data. Each of the pitch contours in the raw data is quantitatively represented by a fairly large number (350–650) of pitch-period samples. Two problems arise if this data is used directly in the speaker-recognition tests. First, each pitch contour is not represented by the same number of samples. Second, in order that Eq. 2 has a solution, either B or W must be nonsingular. For W to be nonsingular, the number of pitch samples should not exceed the total number of known utterances of all speakers. Similarly, for B to be nonsingular, the number of pitch samples should be smaller than the number of speakers. The first problem can be solved by resampling the pitch contours at a uniform rate. In the second case, the objective is to reduce the amount of data without losing significant information concerning the interspeaker differences.

The steps to reduce the number of pitch samples were taken in two stages. First, a pitch contour (duration normalized to 2 sec) was divided into 40 contiguous

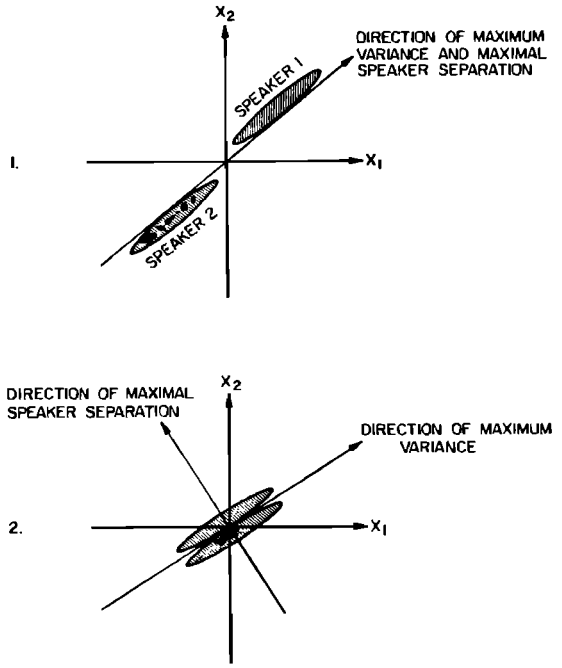

FIG. 3. Directions of maximum speaker separation and maximum variance.

segments, each 50 msec in duration. Each of these segments was then characterized by the average value of the pitch samples in that segment. Let N_k be the number of pitch-period samples in the kth segment and let p_j be the duration of the pitch period at time t_j. Then, the average sample value x_k' characterizing the kth segment is given by

$$x_k' = \frac{1}{N_k} \sum_{j=1}^{N_k} p_j, \tag{17}$$

and $0.05(k-1) \leq t_j < 0.05k$. Each pitch contour was thus represented after the first stage of data reduction by a total of 40 average pitch samples. The above data reduction procedure can also be regarded as a smoothing operation on the pitch contour. The loss of high frequencies accompanying the smoothing operation was not considered serious, since the high-frequency variations in the pitch contours did not reproduce themselves consistently among the different utterances of the same speaker.

Next, the pitch samples x_k' were averaged over the 60 utterances of the 10 speakers. The average values $\bar{x}_k'$ do not contribute to interspeaker differences, and consequently can be subtracted out. The pitch sample $x_k^{(i)}$ characterizing the kth segment of the ith pitch contour, after subtracting the average value $\bar{x}_k'$, is given by

$$x_k^{(i)} = x_k'^{(i)} - \bar{x}_k', \tag{18}$$

where

$$\bar{x}_k' = \frac{1}{60} \sum_{i=1}^{60} x_k'^{(i)} = \langle x_k'^{(i)} \rangle_i. \tag{19}$$

In the second stage of data reduction, the pitch data was further compressed by transformation to the Karhunen-Loève (K–L) coordinate system.[19] The K–L coordinate system provides an optimum representation of a set of N-dimensional vectors by another set of vectors of a lower dimensionality. Consider a set of N-dimensional vectors $\{\mathbf{x}^{(i)}\}$. We approximate each $\mathbf{x}^{(i)}$ by a K-dimensional vector $\mathbf{y}^{(i)}$ $(K<N)$ so that the mean square error E, defined as

$$E \equiv \langle \|\mathbf{x}^{(i)} - A\mathbf{y}^{(i)}\|^2 \rangle_i, \tag{20}$$

is minimum, where A is a $N \times K$ matrix. It can be shown[20] that E is minimum if the columns of A are chosen to be the eigenvectors corresponding to the K largest eigenvalues of the equation

$$R_x \psi = \mu \psi \tag{21}$$

and

$$\mathbf{y}^{(i)} = A^t \mathbf{x}^{(i)}, \tag{22}$$

where R_x is the covariance matrix of the vectors $\{\mathbf{x}^{(i)}\}$ and is defined as

$$R_x \equiv \langle \mathbf{x}^{(i)} [\mathbf{x}^{(i)}]^t \rangle_i. \tag{23}$$

The error E is the sum of the remaining $N-K$ eigenvalues, i.e.,

$$E = \sum_{j=K+1}^{N} \mu_j, \tag{24}$$

where $\mu_1, \mu_2, \cdots, \mu_N$ are the eigenvalues of R_x arranged in order of decreasing magnitude.

One may rightly ask at this point if there is any assurance that, in the process of data reduction, information concerning interspeaker differences has not been lost. After all, the error signal discarded in the process of reducing dimensionality might be significant for discriminating one speaker from another. There is certainly no assurance that this is not the case. However, if the variance of the input data consists mostly of interspeaker variance, the data reduction procedure outlined above will not result in any significant loss of information concerning speakers. This point is further illustrated for the two dimensional case in Fig. 3. In the first case, the direction of maximum variance coincides with the direction of maximum interspeaker variance and the above data reduction procedure will not result in loss of information. This is not true in the second case, where the direction of maximum speaker separation is orthogonal to the direction of maximum variance.

To determine whether any speaker-dependent information was lost in the data reduction process, the vectors $\{\mathbf{x}^{(i)}\}$ were transformed into the vectors $\{\mathbf{y}^{(i)}\}$ with the same number of dimensions $(K=N=40)$. The 40 components of $\mathbf{y}^{(i)}$ were next partitioned into four groups of 10 components each; the first group included the first 10 components, the second group included the next 10 components, and so on. The

effectiveness of each group for speaker discrimination was obtained by computing the "divergence," which is a measure of separability of speakers in the data. The divergence is defined as [21,22]

$$H \equiv \langle [\bar{\mathbf{y}}^{(i)} - \bar{\mathbf{y}}^{(j)}]^t W^{-1} [\bar{\mathbf{y}}^{(i)} - \bar{\mathbf{y}}^{(j)}] \rangle_{i,j}, \quad (25)$$

where $\bar{\mathbf{y}}^{(i)}$ is the mean vector for speaker i and equals $\langle \mathbf{y}_\alpha^{(i)} \rangle_\alpha$, $\mathbf{y}_\alpha^{(i)}$ is the 10-dimensional vector representing the αth utterance of the ith speaker, obtained by regrouping of the transformed data, and W, the pooled intraspeaker covariance matrix, is given by

$$W = \langle [\mathbf{y}_\alpha^{(i)} - \bar{\mathbf{y}}^{(i)}][\mathbf{y}_\alpha^{(i)} - \bar{\mathbf{y}}^{(i)}]^t \rangle_{\alpha,i}.$$

Equation 25 can be rewritten as

$$H = \text{Trace } W^{-1} \langle [\bar{\mathbf{y}}^{(i)} - \bar{\mathbf{y}}^{(j)}][\bar{\mathbf{y}}^{(i)} - \bar{\mathbf{y}}^{(j)}]^t \rangle_{i,j} \quad (26)$$
$$= \text{Trace } W^{-1} B,$$

where B is the interspeaker covariance matrix. The divergence was computed for each of the four groups, the first group consisting of the first 10 components in the K–L coordinate system, the second group consisting of the next 10 components, and so on as described above. The results are shown in Table I. The table shows the divergence, the percent interspeaker variance, and the percent total variance contributed by the four groups. The first group contributes 98.5% of the total variance and 98.1% of the interspeaker variance in the input data. The above results indicate that the input pitch data can be effectively represented in 10 or possibly 20 dimensions.

TABLE I. The divergence, the percent interspeaker variance, and the percent total variance contributed by the four groups of Karhunen–Loève components.

Group	Divergence	Percent interspeaker variance	Percent total variance
1	44.9	89.1	98.5
2	3.5	6.9	1.
3	1.2	2.4	0.25
4	0.8	1.6	0.04

IV. RESULTS OF SPEAKER RECOGNITION TESTS

A. Speaker Recognition Based on the First Group of 10 Karhunen–Loève Components

In these tests, the 10-dimensional vectors $\{\mathbf{y}_\alpha^{(i)}\}$, obtained from the first 10 K–L components, were used to classify the speakers. The classification procedure was carried out as follows:

(1) The pooled intraspeaker covariance matrix W and the interspeaker covariance matrix B, defined in Eqs. 3 and 4, respectively, were calculated. The data in the design set only was used to compute B and W.

(2) The eigenvalues and the eigenvectors of the characteristic equation $B\mathbf{u} = \lambda W\mathbf{u}$ were obtained. One typical set of eigenvalues is listed in Table II. The first six eigenvalues account for all but 2.3% of the interspeaker variance. Since an eigenvalue is a measure of the effectiveness of that particular coordinate in the transformed space for distinguishing between speakers, these results show that only a few of the 10 dimensions are significant for discriminating one speaker from another.

TABLE II. Eigenvalues of the characteristic equation $B\mathbf{u}_n = \lambda W\mathbf{u}_n$.

n	λ_n
1	24.60
2	14.13
3	9.82
4	2.43
5	1.28
6	0.99
7	0.14
8	0.09
9	0.01
10	0.00

(3) The transformation T_M was formed with the eigenvectors $\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_M$ as rows. The recognition procedure was repeated for every value of M between 1 and 10 to determine the optimum number of dimensions.

(4) The mean vectors of the speakers were transformed according to the equation

$$\mathbf{z}^{(i)} = T_M \bar{\mathbf{y}}^{(i)}, \quad (27)$$

where $\bar{\mathbf{y}}^{(i)}$ is the mean vector of the ith speaker in the original space, and $\bar{\mathbf{z}}^{(i)}$ is the mean vector of the ith speaker in the transformed space.

(5) The vector $\mathbf{y}$ to be classified was also transformed according to the equation

$$\mathbf{z} = T_M \mathbf{y}, \quad (28)$$

where $\mathbf{z}$ is the transformed vector. Next, the Euclidean distance between $\mathbf{z}$ and each of the vectors $\bar{\mathbf{z}}^{(i)}$ was computed. The distance between $\mathbf{z}$ and the mean vector of the ith speaker is given by

$$d_i = \|\mathbf{z} - \bar{\mathbf{z}}^{(i)}\|. \quad (29)$$

The vector $\mathbf{z}$ was finally assigned to the speaker for which d_i was minimum.

The percentage of correct identifications increased from 38% for $M=1$ to 93% for $M=8$. There was no further improvement when M was increased to 10. This could be expected in view of the small values of the last two eigenvalues in Table II. There were four misidentifications out of a total of 60 test utterances.

B. Speaker Recognition Based on the Remaining Groups of K–L Components

In order to determine the usefulness of the remaining K–L components for speaker recognition, these components were used in groups of 10 each in the speaker-

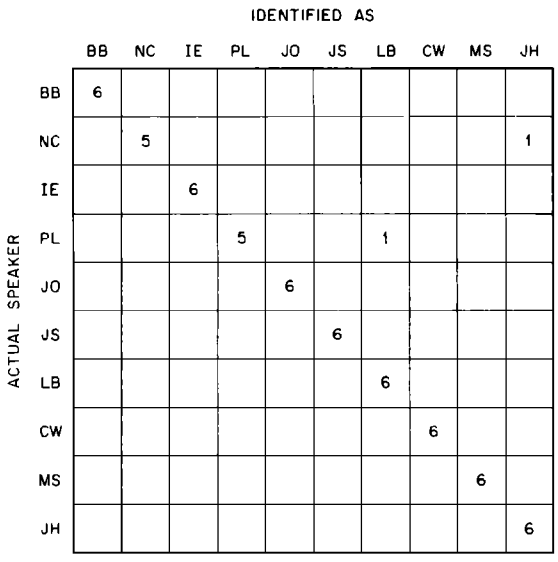

FIG. 4. Confusion matrix of the speakers for the first 20 K–L coefficients.

recognition tests. The procedure was identical to the one described above for the first 10 K–L components. The results of these tests are summarized in Table III. These results indicated that the second group of 10 K–L components could be used to improve the accuracy of identification beyond the 93% obtained on the basis of the first 10 components. The other 20 components contained very little information which could be useful for speaker recognition, confirming the conclusion arrived at earlier on the basis of divergence in Sec. III.

C. Speaker Recognition Based on Combining the First Two Groups of K–L Components

The first two groups of K–L components (1–10 and 11–20) were combined to provide a 20-dimensional representation of the pitch contours. Let $T_M^{(1)}$ and $T_M^{(2)}$ be the linear transformations for the first and second groups of K–L components, respectively. A combined transformation T,

$$T=\begin{bmatrix} T_M^{(1)} & 0 \\ 0 & T_M^{(2)} \end{bmatrix}, \tag{30}$$

where 0 is an $M\times 10$ null matrix (all elements zero),

TABLE III. Percentage of correct identifications for the four groups of Karhunen–Loève components.

Group	Percentage of correct identifications
1	93
2	57
3	20
4	11

and T an $2M\times 20$ matrix, was then obtained. The distance d_i was calculated as before for every speaker and the unknown vector assigned to the speaker corresponding to the closest mean vector.

The classification results based on the above procedure are shown in the confusion matrix of Fig. 4. The results indicate improved recognition, the overall percentage of correct identifications being 97%.

D. Use of Duration for Speaker Recognition

The pitch data used so far in the speaker-recognition tests was based on the utterances normalized to have the same duration. The duration of an utterance itself could, however, be used as an additional independent measure for speaker recognition.

The durations of the utterances were normalized so that the intraspeaker variance of the durations in the design set was unity. A new distance d_i' between the utterance to be classified and the mean vector of the ith speaker was computed. The distance d_i' was defined as

$$d_i'\equiv(d_i^2+\eta_i^2)^{\frac{1}{2}}, \tag{31}$$

where η_i is the difference between the duration of the utterance to be classified and the average duration for the ith speaker.

As a result of adding the duration information to the pitch information, the percentage of correct identifications increased to 98% (59 correct identifications out of a total of 60 judgments). The only remaining error occurred when the speaker PL was classified once as LB.

E. Confidence Interval for the Percentage of Correct Identifications

The 95% confidence interval for the percentage of correct identifications was calculated on the assumption that the total number of errors was binomially distributed. The number of independent trials was assumed to be equal to the number of test patterns. The 95% confidence interval, when both the duration and the pitch information were used for speaker recognition, was found to be 92% to 99%.

V. PRACTICAL CONSIDERATIONS IN IMPLEMENTING THE AUTOMATIC SPEAKER RECOGNITION SCHEME

The most complex operation in the method is the measurement of pitch. In the present simulation, pitch analysis consumed most of the computing time. However, the method used for pitch analysis in the present study is not essential for obtaining the pitch data for an automatic speaker recognition system. An important feature of this method of pitch analysis is its ability to provide period-by-period measure of the pitch period. However, as a result of this study, we found that the period-by-period variations of pitch were not a

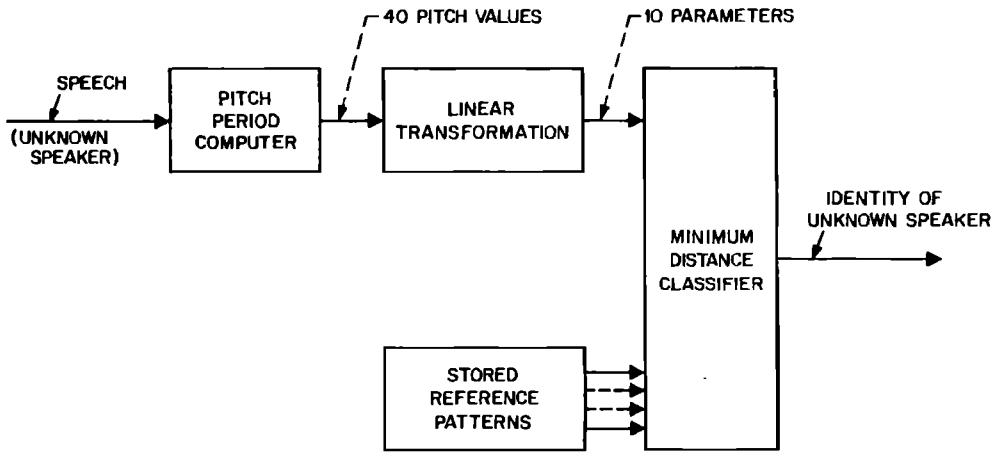

FIG. 5. Block diagram of the automatic speaker-recognition scheme.

characteristic of an individual speaker. Consequently, simpler pitch analysis methods,[23] which operate nearly in real time, can be used, providing greater economy as well as computational speed. As an additional simplification, it is possible to combine the transformations used for data reduction and for maximizing the ratio of interspeaker to intraspeaker variance into one 10×40 matrix transformation. The 10-dimensional vector can then be compared with the stored mean vectors of the different speakers and assigned to the "nearest" speaker. The linear transformation and the classification procedure are functionally very simple and can be implemented in an economical manner. Ideally, a new linear transformation should be obtained whenever a new speaker is added to the population. However, this may not be essential. The updating can be done less frequently. The determination of a new transformation requires that the eigenvalues and the eigenvectors of one 40×40 matrix and four 10×10 matrices be computed. Since these operations need be performed only when a new speaker is added, the computation time needed to identify an unknown speaker will not be affected.

A block diagram indicating the various stages of signal processing required for an automatic speaker-recognition system is shown in Fig. 5.

VI. SUMMARY OF OTHER SPEAKER-RECOGNITION PROCEDURES

The data reduction and recognition procedures described in Sec. III were obtained after testing several other recognition procedures. These other procedures and their results are summarized in this section in order to compare the merits of the different procedures.

A. Minimum-Distance Classification Procedure

The minimum-distance decision rule is one of the least sophisticated of recognition procedures. However, it is widely used due to its computational simplicity; only the mean vector for each class need be determined in the design stage. To classify an unknown vector, its distance from the mean vector of each class is determined. The unknown vector is assigned to the class whose mean vector is closest to it. Thus the decision rule is based on the distance d_i defined by

$$d_i=\|\mathbf{x}-\bar{\mathbf{x}}^{(i)}\|, \tag{32}$$

where $\mathbf{x}$ is the vector to be classified, and $\bar{\mathbf{x}}^{(i)}$ is the mean vector of the ith class. The vector $\mathbf{x}$ is assigned to class j if $d_j<d_i$ for every $i\neq j$.

The above rule was used to classify the speakers based on the 40-dimensional vectors $\{\mathbf{x}_\alpha^{(i)}\}$ representing the means of the 40 consecutive segments of the pitch contours (see Sec. III). The number of correct identifications was 41 out of a total of 60 judgments, representing an accuracy of 68%.

It may be pointed out here that the above procedure differs from the one outlined in Sec. III in that no transformation is used before the distances are computed. Thus, the linear transformation is responsible for reducing the misidentification rate from 32% to 3%.

B. Recognition Procedure Based on Cross Correlation

In this case, the unknown vector was classified by cross correlating the unknown vector with the reference vectors of the different speakers. The unknown vector was assigned to the speaker whose reference vector had the highest correlation. The reference vector for each speaker was obtained as follows:

Let $\mathbf{r}^{(i)}$ be the reference vector for the ith speaker, let $\mathbf{x}_\alpha^{(i)}$ be the vector representing the αth utterance of the ith speaker in the design set, and let $\mathbf{x}$ be an unknown vector of the test set. The cross correlation ρ_i between the unknown vector $\mathbf{x}$ and the reference vector $\mathbf{r}^{(i)}$ of the ith speaker is defined as

$$\rho_i\equiv[\mathbf{r}^{(i)}]^t\mathbf{x}. \tag{33}$$

The reference vector $\mathbf{r}^{(i)}$ was chosen so that $\|\mathbf{r}^{(i)}\|=1$, and the average cross correlation between $\mathbf{r}^{(i)}$ and the vectors of the ith speaker in the design set was maximum. The average cross correlation $\bar{\rho}_i$ for the ith speaker is given by

$$\bar{\rho}_i=\langle[\mathbf{r}^{(i)}]^t\mathbf{x}_\alpha^{(i)}\rangle_\alpha. \tag{34}$$

The vector $\mathbf{r}^{(i)}$, which maximizes $\bar{\rho}_i$ under the constraint $\|\mathbf{r}^{(i)}\|=1$, is given by

$$\mathbf{r}^{(i)}=\bar{\mathbf{x}}^{(i)}/\|\bar{\mathbf{x}}^{(i)}\|, \tag{35}$$

where $\bar{\mathbf{x}}^{(i)}=\langle\mathbf{x}_\alpha^{(i)}\rangle_\alpha$.

The 40-dimensional vectors $\{\mathbf{x}_\alpha^{(i)}\}$ were again used to classify the utterances of the test set. The number of correct identifications was 42, representing a recognition accuracy of 70%. The cross-correlation procedure thus offers no significant advantage over the minimum-distance classification method.

C. Recognition Procedure Based on Moments of Pitch-Period Distribution

A different method of representing the pitch data based on histogram analysis was tried. It is possible that the first-order probability distribution of pitch is speaker dependent and consequently could be used to recognize speakers. In order to test this hypothesis, pitch histograms giving the frequency of occurrence of a pitch period in the pitch contour were obtained. The entire pitch range from 0 to 10 msec was divided into 50 equal parts. The number of times a pitch value in a given pitch range occurred in a particular utterance was counted automatically on the computer. This process was repeated for the 60 utterances of the 10 speakers.

In order to determine whether the gross characteristics of the histograms could be used for speaker recognition, the first four central moments of each of the 60 utterances were calculated. The four moments were then subjected to the multidimensional linear discriminant analysis outlined in Sec. III. The results of speaker-recognition tests showed that speakers could be identified with an accuracy of 78% on the basis of the first four moments. This error rate is approximately two-thirds of the error rate obtained by the cross-correlation method but still seven times higher than that obtained by using the temporal information in the pitch contour. This histogram analysis does not utilize time-dependent information in the pitch contours. For example, a pitch contour could be reversed in time without altering the histograms. Thus, considerably lower error rates are achieved by utilizing the time-dependent information in the pitch contours.

VII. SUMMARY AND CONCLUSIONS

The main objective of the present investigation was to determine the usefulness of pitch contours for automatic speaker recognition. The results of the speaker-recognition tests described in Sec. IV indicate that pitch contours can be used effectively for speaker recognition. Although these tests were carried out on a population of only 10 speakers, the results of these tests could be expected to be valid for a much larger population; the average duration of the pitch periods for the 10 speakers ranged only between 4.3 and 5.6 milliseconds whereas the entire pitch range in speech extends from 2.5–15 msec.

The most efficient recognition procedure was found to be one based on multidimensional linear discriminant analysis. This approach allowed us to single out and discard those features in the data which were not significant for speaker recognition. The error in identifying a speaker was found to be only 2%.

The Karhunen–Loève coordinate system was found to be an efficient method of compressing the data in the present case. The original data, where every pitch contour was represented by approximately 500 pitch samples, was effectively compressed to two sets of 10-dimensional vectors. The reduction in the dimensionality of the data permitted the use of sophisticated statistical techniques without running into computational and storage problems on the digital computer.

The high-frequency information in the pitch contours did not reproduce itself consistently among the utterances of the same speaker. The speaking rate varied considerably even among the utterances of the same speaker. The variability due to different speaking rates was reduced by adjusting every utterance to have the same duration. A part of the intraspeaker variance in the high-frequency information was caused by the residual variability which could not be eliminated by normalization of the durations alone. It is possible that better time normalization of the utterances could lead to smaller intraspeaker variance in the high-frequency information; this in turn could provide more reliable speaker identification.

No attempts were made to find a set of optimum utterances for speaker recognition. The particular utterance used in the study had the advantage of being completely voiced. Utterances consisting of both voiced and unvoiced segments may be advantageous for time normalization. The location of unvoiced segments could be used to align the time coordinates of different utterances. There was some evidence that utterances involving considerable movement of the articulators might be better for speaker recognition.

Previous speaker-recognition methods were based on the short-time spectral characteristics of speech. The investigation reported in this paper provides evidence that pitch contours could also be employed successfully for automatic speaker recognition. The choice of speech characteristics suitable for speaker recognition, however, need not be restricted to the short-time spectrum and pitch contours alone. One important set of characteristics which has not been utilized so far for automatic speaker recognition is represented in the temporal

variations of formant frequencies. As in the case of pitch contours, formant frequencies have the advantage of not being affected by the frequency characteristics of the transmission system, the level at which the speaker talks, and the distance between the microphone and the speaker. The classification and data reduction procedures outlined in Sec. III could be used directly to test the effectiveness of formant frequencies for speaker recognition.

ACKNOWLEDGMENTS

The author wishes to express his appreciation to Mischa Schwartz, the author's thesis advisor, for his guidance and valuable advice in all phases of this research. The author is also indebted to M. R. Schroeder, P. B. Denes, and M. M. Sondhi for many helpful discussions and comments.

* The research reported in this paper formed part of a doctoral dissertation submitted by the author to the Dept. of Elect. Eng., Polytech. Inst. of Brooklyn, New York (June 1968).

† Presented in part at the 76th Meeting of the Acoust. Soc. Amer., Cleveland, Ohio, 18–22 Nov. 1968 (J. Acoust. Soc. Amer. **45**, 309 (A) (1969)].

[1] F. McGehee, "The Reliability of the Identification of Human Voice," J. Gen. Psychol. **17**, 249–271 (1937).

[2] I. Pollack, J. M. Pickett, and W. H. Sumby, "On the Identification of Speakers by Voice," J. Acoust. Soc. Amer. **26**, 403–406 (1954).

[3] P. D. Bricker and S. Pruzansky, "Effects of Stimulus Content and Duration on Talker Identification," J. Acoust. Soc. Amer. **40**, 1441–1449 (1966).

[4] S. Pruzansky, "Pattern-Matching Procedure for Automatic Talker Recognition," J. Acoust. Soc. Amer. **35**, 354–358 (1963).

[5] W. Hargreaves and J. A. Starkweather, "Recognition of Speaker Identity," Language and Speech **6**, 63–67 (1963).

[6] K. P. Li, J. E. Dammann, and W. D. Chapman, "Experimental Studies in Speaker Verification Using an Adaptive System," J. Acoust. Soc. Amer. **40**, 966–978 (1966).

[7] 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. **50**, 1427–1454 (1971).

[8] J. E. Luck, "Automatic Speaker Verification Using Cepstral Measurements," J. Acoust. Soc. Amer. **46**, 1026–1032 (1969).

[9] S. K. Das and W. S. Mohn, "Pattern Recognition in Speaker Verification," AFIPS Conf. Proc., Fall Joint Computer Conf. **35**, 721–732 (1969).

[10] G. R. Doddington, "A New Method of Speaker Verification," J. Acoust. Soc. Amer. **49**, 139(A) (1971).

[11] B. S. Atal, "Automatic Speaker Recognition Based on Pitch Contours," Ph.D. thesis, Polytech. Inst. of Brooklyn (June 1968), pp. 22–43.

[12] B. S. Atal, "Period-by-Period Pitch Analysis of Speech," (to be published).

[13] Complete pitch data for the 60 utterances are given in Ref. 11.

[14] G. S. Sebestyen, *Decision-Making Processes in Pattern Recognition* (MacMillan, New York, 1962), pp. 8–53.

[15] R. A. Fisher, "The Use of Multiple Measurements in Taxonomic Problems," in *Contributions to Mathematical Statistics* (Wiley, New York, 1950), pp. 32.179–32.188.

[16] Ref. 11, pp. 119–122. See also Ref. 14, pp. 40–43.

[17] S. S. Kuo, *Numerical Methods and Computers* (Addison-Wesley, Reading, Mass., 1965), pp. 199–206.

[18] The matrix C is called the modal matrix of W.

[19] S. Watanabe, "Karhunen-Loève Expansion and Factor Analysis," in *Transactions of the Fourth Prague Conference on Information Theory, Statistical Decision, Functions, Random Processes* (Academia Publishing House, Czechoslovak Academy of Science, Prague, 1967), pp. 635–660.

[20] H. P. Kramer and M. V. Mathews, "A Linear Coding for Transmitting a Set of Correlated Signals," IRE Trans. Information Theory **IT-2**, No. 3, 41–46 (1956).

[21] S. Kullback, *Information Theory and Statistics* (Wiley, New York, 1959), p. 190.

[22] T. Marill and D. M. Green, "On the Effectiveness of Receptors in Recognition Systems," IEEE Trans. Information Theory **IT-9**, No. 1, 11–17 (1963).

[23] B. Gold and L. Rabiner, "Parallel Processing Techniques for Estimating Pitch Periods of Speech in the Time Domain," J. Acoust. Soc. Amer. **46**, 442–448 (1969).