A Scheme for Speech Processing in Automatic Speaker Verification

S. K. DAS
W. S. MOHN
IBM Corporation
Research Triangle Park, N. C. 27709

Abstract

Experiments investigating adaptive pattern recognition in automatic speaker verification are reported. A binary decision confirming or rejecting a speaker's purported identity is required. The experiments involve 7000 phrase length utterances of 118 speakers. An average misclassification rate of one percent with a "no decision" rate of ten percent is obtained. Other experiments indicate that the utterances used for training purposes should preferably be collected over a relatively long period of time.

Manuscript received July 15, 1970.

I. Introduction

The problem of speaker verification continues to be a challenge to many investigators [1], [2], [4]. In this problem, the identity of a person who utters a certain phrase is confirmed from an analysis of the phrase and some knowledge of the speaker's speech characteristics obtained from earlier utterances of the same phrase. The person is referred to as a "real" speaker. At the same time, the system should be able to reject other speakers who might falsely claim to be the real speaker. Thus, they are referred to as "impostors." A real speaker and the impostors represent the two classes that the verification system is designed to separate.

Implementation of a speaker verification system involves a combination of speech processing techniques and pattern recognition methodology. This will be clear from Fig. 1, which depicts the proposed system. Utterances of a suitable phrase are first tape recorded by all of the speakers. These analog utterances are then digitized by using a vocoder type of speech analyzer (Section II). These are the preliminary steps of speech processing and are carried out in hardware. The digitized data are stored on digital tapes and form the input to a software system. The first stage of this software system serves to segment the utterances (Section III). This is essentially a pattern recognition scheme in which various events in the utterances are identified. The identified events are used to align the utterances properly so that appropriate features can be extracted from the segmented data (Section IV). This is a further stage of speech processing and is expected to yield feature sets suitable for speaker discrimination. These feature sets are used in a pattern recognition task for creating one reference profile for each real speaker or, in a different operational mode, for making a verification decision (Section V).

It is interesting to compare the tasks of the segmentation and the feature extraction process. The first one is designed to detect corresponding events from various utterances of all speakers. Thus, it places emphasis on phenomena that are similar among the speakers, but different among the events constituting the speech sample. On the other hand, the feature extraction procedure is designed to obtain features which distinguish speakers from one another. Thus, it places emphasis on phenomena that are different among the speakers, but are similar in terms of the speech events.

This paper describes principally the procedures followed in data collection, digitization, segmentation, and feature extraction. The techniques for creating a reference profile for a speaker and for eventual decision making are dealt with briefly since they have been described in detail elsewhere [1]. The structure of the verification system is such that it is fully automatic after its implementation. This feature is essential in order that the system can be feasible in practical applications and in order that a large-scale experiment such as the one reported here for performance validation may be under-

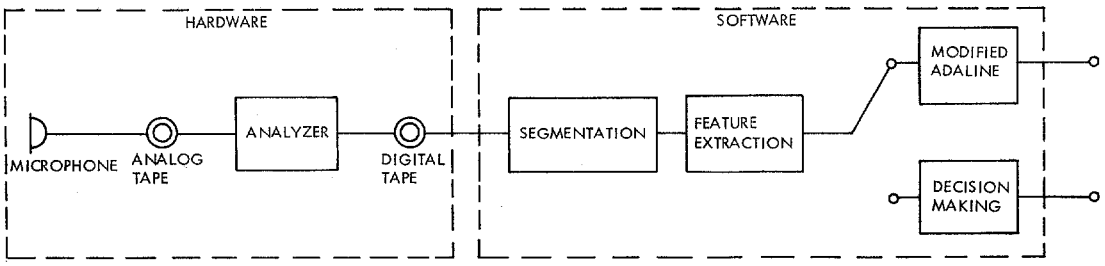


Fig. 1. Experimental setup.

taken. The last section of the paper includes some comments about the experiments and some observations regarding possible future work.

II. Data Collection

One of the significant factors that distinguishes this work from earlier reported experiments in the same field by other research workers is the size of the data base [2]–[4]. In speaker verification, large amounts of data are required for several reasons. In the case of each real speaker, sufficient data must be available to "train" the verification system, and an independent set of utterances must be available to later test the system. In the case of impostors, many speakers must be available to test the system's sensitivity to nonreals, and in some cases to "train" the system in the first place. Complications arise in the collection of sufficient data since each speaker must be sampled at a number of different times to gain a representative sample of his speech variability. In other words, all of the training utterances cannot be gathered at one time if accurate verification of later utterances is desired. It has been found that utterances should be collected over a period of several weeks, ideally being spread uniformly over that period. It is difficult to gather such data in a controlled manner so one is forced to schedule recording sessions in which several utterances are collected at one time.

An easy way to organize a recording session would be to have the speaker repeat the chosen verification phrase a number of times, one right after the other. This violates the requirement that the utterances must be representative of isolated examples from the speaker as would be encountered in practical use. Therefore, some task must be inserted between each utterance to "neutralize" the speaker. The data collection procedure about to be described followed these guidelines.

The five phrases listed in Table I were selected for the data collection phase. They were selected according to the following criteria: 1) each must be made up of words familiar to the speakers; 2) each should begin with a high-energy sound to cause reliable starting of the digitizing hardware; 3) each should contain points that allow reliable segmentation of key phonetic events; and 4) each should also contain a maximum number of independent phonological events carrying speaker-specific information. A randomized ordering containing each phrase exactly ten times was created. At a given recording session, a speaker was asked to enter an acoustically treated chamber which effectively isolated the speaker acoustically from the laboratory and reduced sound reflection within the chamber. In the chamber, the speaker seated himself at a small table upon which was located a phrase indicator, a digital indicator, and a ten-digit keyboard. The speaker placed a Koss KR2+2 headset and boom-mounted microphone combination on his head. The microphone was connected to the inputs of one track of each of five tape recorders [Ampex 351–2's) to separate phrases by type. The other track of each tape recorder was used to record tones resulting from key depressions of the ten-digit keyboard. The tape recorders were interlocked so that only one could be running at a time. The phrase indicator lights were lit to indicate which phrase should be spoken. The digital indicator was simply a counter to indicate how far through the list of 50 phrases the recording has progressed.

TABLE I

Five Experimental Phrases

1) Check Intermediate Allowance
2) Check Numeric Interrogation
3) Check Available Terminals
4) Check Allowable Clearance
5) Check Preliminary Correlation

Note: All phrases recorded; only 3) digitized.

During operation, a tone coded label was recorded on each tape recorder to identify each speaker. Next, the experimenter started the tape recorders in the sequence indicated by the randomization list. As a tape recorder was started, the digital counter was incremented, the speaker keyed in the counter reading and then uttered the phrase. Generally, the speaker could not anticipate which phrase was coming next, and there was a guaranteed pause between utterances caused by the phrase-labeling keyboard task and the time during which the experimenter outside the chamber switched from one tape recorder to another. This procedure resulted in the production of sets of tapes, each containing only a single phrase, thereby simplifying eventual experimental processing.

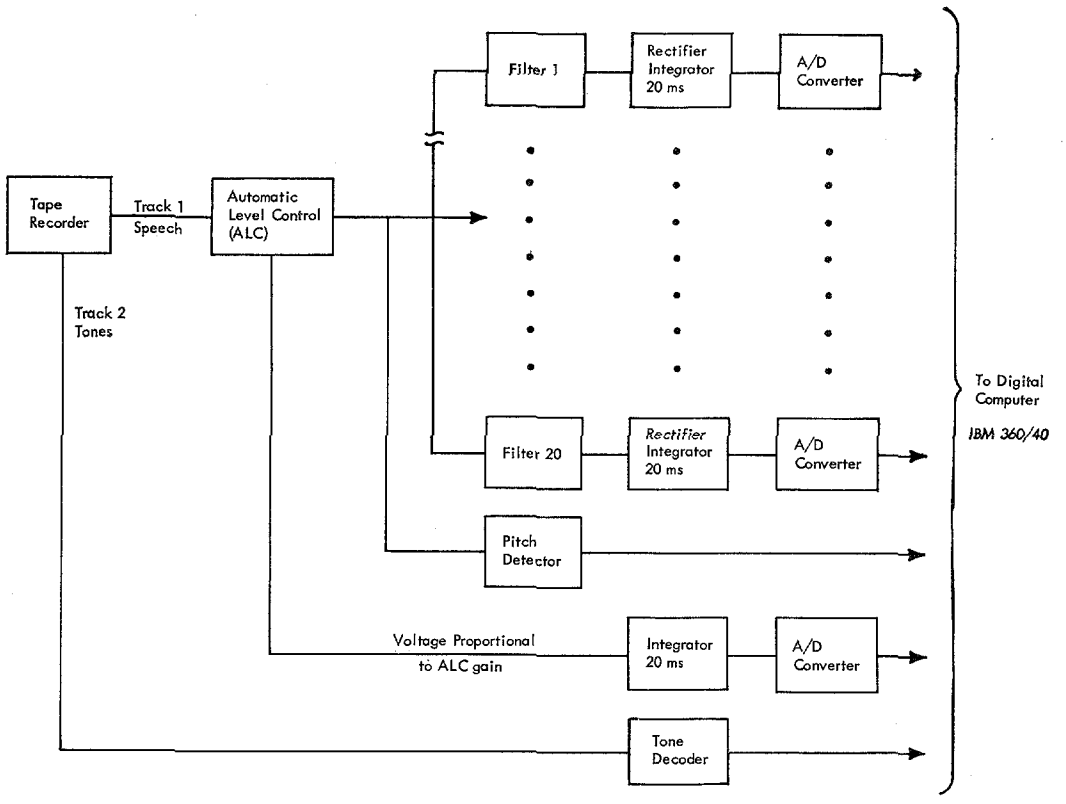


Fig. 2. Automatic digitizing system.

After the experimenter had obtained 50 utterances, the speaker was asked to leave the chamber and rest for five to ten minutes. Following this pause, the speaker reentered the chamber and the same list of phrases was recorded, in the same sequence as before. This full set of 100 utterances, 20 for each phrase, constituted a single recording session. This procedure allowed the collection of a great many utterances while maximizing utterance independence under the given time constraints and not occupying too much of each speaker's time.

The original plan called for a recording period covering five weeks using speakers who were male engineers and scientists in the laboratory. During this time, the speakers, who were arbitrarily designated as "reals," were expected to attend one recording session a week. Meanwhile, each of the speakers designated "impostor" was asked to make one recording session. Thus, each real was to produce 100 utterances of each phrase while each impostor produced only 20. Fewer were needed from each impostor since his utterances were eventually to be used for either training or testing but not both. Initially, 58 people were designated as reals and 69 as impostors. For several reasons the recording period exceeded five weeks and not all of the speakers produced a full set of utterances. This was the result of speakers being transferred, absent, or otherwise unable to attend their recording sessions on time. This resulted in only 50 usable reals and 68 usable impostors. A total of 7180 utterances of each phrase was collected by this procedure. However, the useful number was somewhat less than this for reasons about to be discussed.

TABLE II

Filter-Bank Specifications

Filter Number	Center Frequency (Hz)	±6.0 dB Bandwith (Hz)
1	188	250
2	459	250
3	715	250
4	969	250
5	1220	250
6	1472	250
7	1725	250
8	1975	250
9	2225	250
10	2475	250
11	2725	250
12	2991	290
13	3300	330
14	3659	390
15	4083	460
16	4586	550
17	5194	670
18	5954	860
19	6932	1110
20	8203	1450

The digitizing system is shown in Fig. 2. The speech analyzer consists of 20 bandpass filters covering the range of center frequencies from 188 to 8203 Hz. The center frequency and the bandwidth of each filter are listed in Table II. The output of each filter is full-wave rectified and passed to a resettable integrator with a 20-ms integration time. The value of each integral is logarithmically coded as a four-bit value spanning a 30-dB range. In addi-

tion, an automatic level control (ALC) acts upon the input to keep the signal within the dynamic range of the A/D converters. The average value of the gain of the ALC is also A/D converted every 20 ms. Finally, a separate fundamental frequency (F_0) measurement circuit quantizes and ditigizes the measured values of F_0 every 20 ms. The analyzer is described in more detail elsewhere [1], [5].

In addition to the speech analyzer, the digitizing system as a whole consisted of three parts: a tape recorder, a device for converting the tones on a second analog tape track to label information, and a digital computer system. The digitizing program performed the following functions. It first detected the speaker's identity from the initial tone sequence. Next, the phrase number was detected and the analyzer began digitizing the speech track following the receipt of the last tone. Digitizing continued until either 6 s had elapsed or the tone sequence for the following utterance began. The 6-s figure made the digitized record length reasonably short, yet while truncating very few utterances. Only one phrase, "check available terminals" has been digitized to date.

After the data were digitized, several steps were taken to insure that only valid examples of the phrase were kept for further analysis. This reduction process is summarized in Table III. (The set of analog utterances was 87 less than the planned number because speakers sometimes did not appear for recording sessions or did not provide the proper number of utterances at a recording session.) Occasional analog recordings were not digitized due to tone decoding error or digitizing program error. Some digitized utterances were removed if the person monitoring the digitizing detected either an incorrect or improperly spoken phrase, or if an utterance was digitized with the wrong label.

A program was written to eliminate silence before "check" and after "terminals." Utterances were eliminated if, after the removal of leading and trailing silence, less than 50 samples remained. It was found experimentally that no one said the full phrase in less than 1 s. This data base of almost 7000 utterances was placed on disk storage so that the next steps of segmentation and feature extraction could be performed efficiently. At this point, of the 7093 utterances of the phrase which had actually been uttered by our speakers, about 98 percent (6926) were still available and thought to be valid, properly labeled examples of "check available terminals."

TABLE III

Presegmentation Utterance Rejection

	Number of Utterances
Theoretical original analog tape	7180
Usable original analog tape	7093
Not digitized (tone sequence errors, program errors, etc.)	47
Initial digital tape	7046
Operator detected digitizing errors	59
Unrecoverable labeling errors	25
Phrase too short after silence removal	36
Resulting digitized set	6926

III. Segmentation

The previous section describes the process of collecting and digitizing the speech samples for the verification experiments. Since all the speech samples pertain to a single phrase, generally each digitized utterance may be assumed to be a collection of a known set of phonetic events in a fixed sequence. On the average, it takes about 1.5 s to pronounce the experimental phrase and considerable variations in speaking rate in different parts of different utterances are inevitable. These cause the time intervals elapsed between the various events to be inconsistent from one utterance to the next. Thus, some of these events must first be identified so that approximately identical parts of various utterances can be compared to each other. This process of identification of events is termed "segmentation."

For several reasons, the process of segmentation is one of the most critical steps. The importance of accurately performing this step cannot be overemphasized. Gross recognition errors are likely to occur if incorrectly segmented utterances are used for later processing. Previous investigators have often based their utterance alignment on the event with the maximum amplitude or energy [3], [4]. For some reported experiments, segmentation has been performed manually [6]. The former procedure is not suitable for our experiments due to the complexity of the uttered phrase; the points of maximum amplitude or energy will not be consistent with respect to the various events of interest in many utterances. The manual procedure is unacceptable due to the proposed large scale of our experiments and due to our intention to make the system fully automatic. Thus, it is necessary to adopt the laborious method of studying a number of spectrograms and devising a set of rules to detect the phonetic events of interest. The task is complicated by the broadness of the range of intraspeaker and intra-utterance variations that can be expected in the spectral representation of a given event. The crux of the segmentation scheme lies in taking into account such wide variations and yet not misidentifying any event. Due to such demands, the software programs which are designed to accomplish segmentation often become so involved with various conditional loops and branches that it is impossible to present a detailed account of their operation here. Thus, in this section, only the segmentation philosophy will be outlined; no attempt will be made to describe all of the rules.

Another matter of importance is to determine the generalization capability of the designed segmentation rules. It has been established that this is often an iterative procedure [1] in which alternate stages of designing and testing are conducted. Thus, the rules are first designed by using several utterances; next, the rules are tested for generalization on a comparable number of new utterances. If segmentation is not of desired accuracy, this test set is combined with the original design set to

form the new design set. The rules derived from this design set are then tested on a comparable separate set of utterances. This process is repeated until the generalization is sufficiently accurate. At least the last generalization test should be on a considerable number of utterances from many new speakers whose utterances have never entered into the previous design stages. This should provide some confidence in the accuracy obtained. During each of the preceding design stages, the utterances in the design set must be hand-segmented by inspecting the digital spectrograms and the implicit rules followed in this process must be transformed into a set of software programs. During the generalization test, the utterances selected for this purpose are also hand-segmented in a similar manner and the segmentation points determined by the software programs are compared against the hand-segmented points for acceptability. The criteria for acceptance are entirely subjective. Thus, if an event is of short duration in an utterance, a difference of even one time sample between the hand-segmented and the program-generated points may be objectionable; but, for an event of relatively long duration, slight discrepancies of one or two time samples may be quite acceptable.

It should be clear from the previous discussion that a large amount of effort must be devoted to implementing the segmentation strategy. Some help may be obtained in the above implementation if an interactive display facility (such as an IBM 2250) is used on line. Possibilities of this approach are outlined in Section VI. In the present experiments, the phonetic event corresponding approximately to "v" in "available" is located first. (The events will be given various names for ease of description even though the names may not correspond to these phonetic events precisely in time. The important point to note is that as long as the events are determined satisfactorily in the utterances in their expected locations, the particular names attributed to these events are of little concern.) The only reason for this selection is the conviction arrived at after studying many spectrograms of the test phrase that this is the easiest event to detect with some certainty. After "v" is found, the knowledge of its location is used to determine the events of "available terminals" corresponding roughly to "l," "b," "t," "e," "s," and "m" in that sequence. Further description of the strategy of detecting these events is given in the following paragraphs. No segmentation point in "check" is determined since it is believed that little speaker-dependent information can be extracted from the pronunciation of this word as the first word of the phrase. It is pointed out here that this method of segmentation is subject to criticism. Some probable ways of improving on this approach are outlined in Section VI.

In order to locate "v," it is necessary to scan a digitized utterance from its beginning. Some simple preliminary rules are used first to find a probable location for "v"; then, a more stringent set of rules is applied to verify if it indeed is "v." If it is not "v," the scanning proceeds again.

In the preliminary set of rules, it is helpful to look for sufficient energy in filter 4 to filter 8 (844 to 2100 Hz). This is approximately the region of the second formant which appears relatively strongly at "v." Thus the filter values in this range are compared against an experimentally determined threshold. Among the more precise set of rules for detecting the presence of "v" are the following: 1) the closure before "-ble" must occur within a reasonable time period from the proposed location of "v." The criterion for what is reasonable was determined during hand-segmentation of the design data. 2) The shape of the second formant within the period from "v" to closure before "-ble" must conform to the characteristics normally encountered in the design utterances. Some variations may be expected in this shape. For example, even though the second formant often increases in frequency at the beginning of this interval, occasionally no conclusive rise is detected. The program should be able to handle these variations. 3) The first formant (filter 1 to filter 3) should be sufficiently prominent. This is determined by comparing with a second threshold.

Sometimes a "v" detected by the program is several samples earlier than what is judged to be the proper location of "v." To remedy this situation, an attempt is always made to find a second "v" location after the first one is located using the previously described plan. If a second "v" is found and if it is within a short interval from the first, the second one is taken to be the proper "v" location. As an additional confirmation, the ALC gain waveform is also checked for its typical shape around this point.

The location of "l" in "available" is determined next. The second formant which appeared strongly at "v" continues up to the this point. Thus, it is necessary to track the second formant starting from "v" and determine the location where this formant stops dropping in frequency. This point is called "l."

Next, the events called "b" and "t" are detected. "b" is evidenced by the rise of voiced energy following the b-closure. "t," similarly, is marked by the t-release aspiration following the t-closure. Two programmed functions of the digitized data were written to detect the presence of voicing and frication. The former considered the power in the lowest three filters and the latter considered the power in the upper 12 filters. The power in a set of contiguous filters was calculated by converting the log-encoded filter values produced by the hardware back to linear power scale, summing, and then reconverting to log scale. Thus, the frequency band (defined by the 6-dB points) of 63 to 840 Hz was used for voicing detection and 2100 to 8900 Hz for fricative detection. The latter range extends to somewhat lower frequencies than would normally be needed for pure fricatives but it was used to detect the T-release which was sometimes more D-like than T-like. The power in each of these bands was compared to an appropriate threshold to characterize each sample as V (primarily voicing), F (primarily frication), B (both), or none of these. To handle what unreliability

remained in the F/V-finding program, the rules for "b" and "t" mentioned above were refined as follows.

The "-ble" section was determined to begin at the last continuous set of V samples before the first F samples following the previously determined segmentation point "l" (in "-vail"). The specific sample labeled "b" in this "-ble" section was that showing the sharpest increased in amplitude as evidenced by the sharpest drop in ALC gain. Thus, "b" reflects the b-release. The "t" in "term-" was found in an analogous way.

The sample labeled "e" in "term-" is the first sample of the V section immediately following the sample labeled "t."

Using these F and V functions, the first sample of the fricative ending "terminals" was called "s." The "ending" fricative was defined to be the first fricative following the last voiced sample in the utterance. Fig. 3 describes an example of the operation of these rules.

These rules have been stated in some detail to show how the apparently simple task of detecting closures in the phrase "check available terminals" can become involved when it is desired to find these closures accurately over an essentially unlimited set of talkers. Even these rules were designed only for male speech.

The last segmentation point detected is "m" in "terminals." This is done primarily by noting the behavior of the second formant region (filters 4, 5, and 6 or 844 to 1587 Hz) after "e." At "m," this formant is least prominent. The ALC waveform shape is also taken into consideration in selecting a final location.

As mentioned before, the segmentation design procedure utilizes a number of utterances in an iterative manner. For the present experiments, a total of about 300 utterances from 20 speakers was used at the final stage of the design process. The generalization capability of the segmentation strategy was checked by testing the rules on approximately 300 utterances of 20 new speakers not included in the design process. One or more of the desired events were undetected in approximately ten percent of the test utterances. Suitable techniques for dealing with these utterances may be devised at a later time. For the present experiments, however, these utterances were simply eliminated from further consideration. Thus, the phrase rejection rate was ten percent. Among the remaining utterances of the test set, "b" was detected incorrectly in one utterance and "m" appeared to be wrong in 24 utterances. When "b" was misidentified, it caused errors in the locations of "t," "e," "s," and "m" as well. One predominant source of error in "m" detection seemed to be the misrecognition of "n" in "terminals" as "m." Besides these, no other errors were apparent.

At this stage, the evaluation of the segmentation strategy was complete. Latter processing steps were carried out without any manual intervention. The total set of 6926 utterances was processed under the segmentation strategy outlined in the preceding paragraphs. Except for the utterances automatically flagged due to the phrase rejection procedure described above,

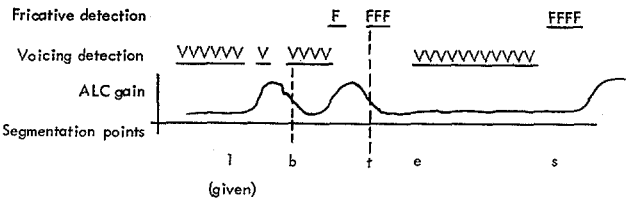


Fig. 3. Example of operation of *l*, *b*, *t*, and *s* finding rules.

all the other utterances segmented in this manner were used as inputs to the next software stage of feature extraction. No attempt was made to disregard the utterances that were segmented incorrectly since that would violate our principle of making the system fully automatic.

IV. Feature Extraction and Evaluation

Most investigators currently agree that the feature extraction step is one of the most crucial in the solution of practical pattern recognition problems. The segmentation effort described in the previous section would be viewed by some as simply a kind of feature extraction. We have chosen to look upon it as a preprocessing step allowing the extraction of more sophisticated features than would be otherwise available.

A total of 405 different features were calculated for each utterance based upon these segmentation points. The features were in general, multilevel discrete approximations to continuous features. Their discrete nature was dictated by the quantization performed by the analysis hardware. The set of features included essentially all that were considered potentially useful. Some features were suggested by various articles in literature [3], [6].

The forms of the various sorts of features will be given below. Reference should be made to Fig. 4(A)–(C) which diagramatically shows all of the features. Only those shaded were eventually used in decision making. (See subsequent paragraphs regarding evaluation procedure.)

Filter Averages

The most common type of feature was the so-called filter average. Each of these was the combination of values of one or more filters for one or more time samples. In each case, the filter values were first converted from a log to a linear scale averaged over the required number of filters and the average finally reconverted to log scale. The number of filters used for the different features ranged from one to all 20. Some features used filters in the regions of the first, second, third, or higher formants by utilizing filters 1 through 4, 5 through 12, 13 and 14, and finally 15 through 20. The choice of "all filters" was meant to give some measure of the total power in that time sample.

The number of time samples involved in a filter average was also one of three choices, short, medium, or long. A short interval was from one to three time samples (20 to

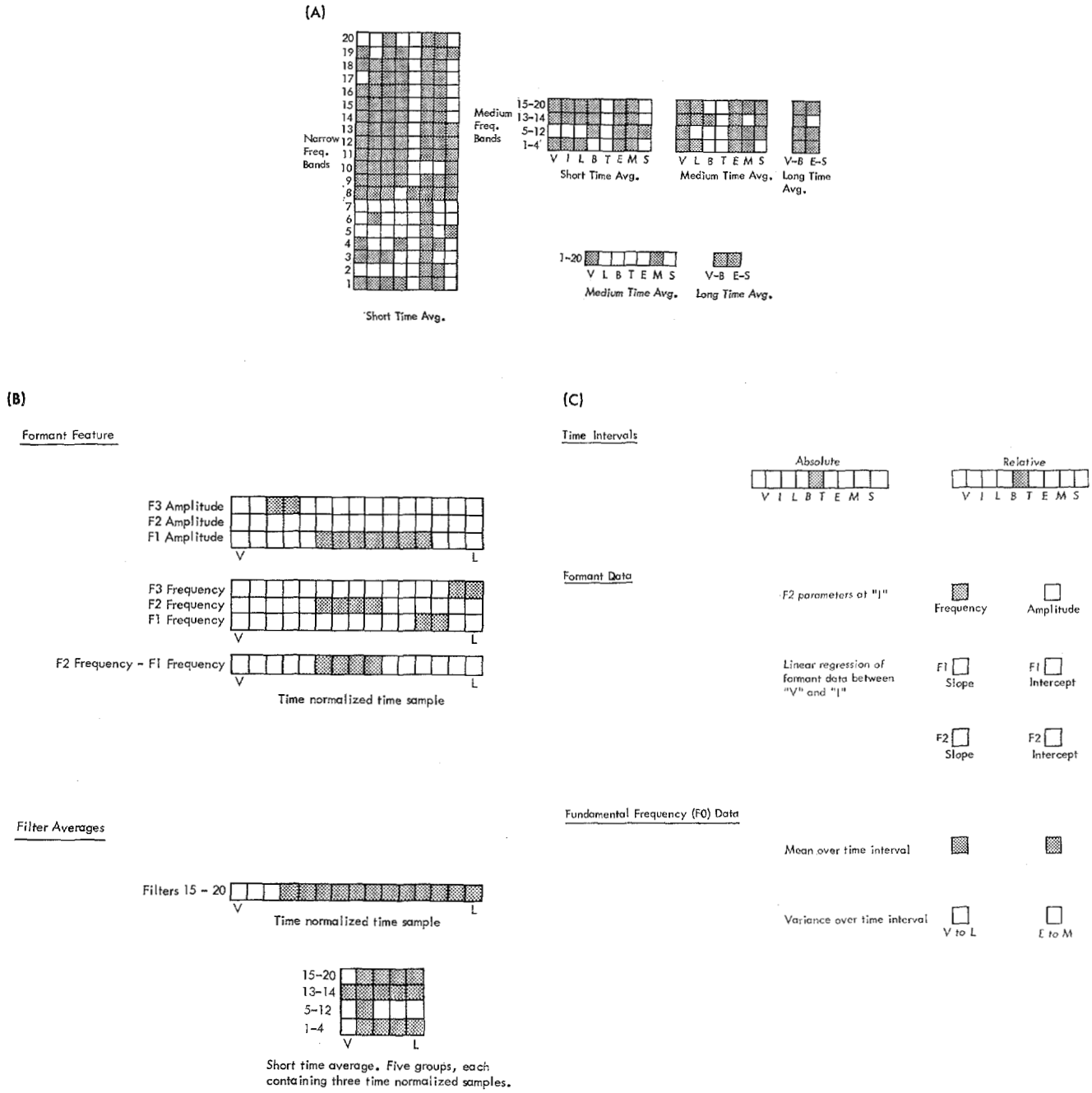


Fig. 4. (A) Features used for speaker verification (filter averages). (Note: Vertical labeling refers to the filters involved in the average. Horizontal labeling refers to the time samples involved in the average. Each block thus defined represents one feature. Only those shaded were retained for later decision making.) (B) Features used for speaker verification (from time-normalized data). (C) Features used for speaker verification (miscellaneous).

60 ms) at a segmentation point. The length varied based upon the particular segmentation point. A medium length interval extended from one segmentation point to the sample immediately preceding the following segmentation point. The long time intervals were between segmentation points "v" and "b," and "e" and "s." This covers all filter averages except those involving time normalized data described below.

Time Normalized Data

Utterances vary in length and simple use of time-frequency-power matrices may not result in sufficiently powerful features. Much of this dependence upon changes in time scale can be eliminated by simple linear time normalization defined as follows. One may take the sampled filter value function represented by a horizontal row in the time-frequency matrix and linearly interpolate

between adjacent sample points. The piecewise linear function can now be resampled with a fixed number of time samples. This was done for the array of samples beginning with "v" and ending with "l." An arbitrary length of 15 samples was chosen for the resampling. Thus a new 20×15 array was formed.

Within the time normalized data two new sets of filter averages were calculated. For one set the array was divided into four frequency bands as before and five equal time intervals of three samples each. The other took only the top frequency band (filters 15 through 20) and considered each new sample independently.

Formant Data

Formant data was extracted in both normal and time normalized data. The following interpolation was performed in the frequency domain in both cases. A four point interpolation formula was used to estimate the power in a hypothetical filter placed between two real filters. The four points were the values produced by the two real filters immediately above and the two immediately below the hypothetical filter in frequency. Thus, the original 20 filters were augmented by 19 interpolated filters to produce a total of 39 upon which formant analysis was performed.

Formant detection operated in two stages. First, each time sample was treated independently, and local peaks were detected in the frequency-interpolated spectrum. These peaks were the input data to the second stage in which three were chosen and labeled the first, second, and third formants, taking into account the normal characteristic of formants not to change greatly in frequency from time sample to time sample. A particular formant at a particular time sample was specified by two parameters: its amplitude measured by the value of the filter output at the peak, and its frequency measured as that of filter numbers 1 through 39. These parameters were included for all 15 time normalized samples for all three formants. In addition, the difference in frequency between F1 and F2 was included in all 15 samples. Other formant features were extracted from the original nontime normalized data also. These were the frequency and amplitude of the second formant peak characterizing the sample labeled "i" in "-vail-" and the intercept and slope information for linear approximations to F1 and F2 between "v" and "i."

Time Difference

Both absolute and relative time differences were included among the features. Time was measured in number of time samples between each adjacent pair of segmentation points. These absolute measures were made relative by normalizing with respect to the number of time samples between "v" and "s." In each case, the time interval from the k-release to "v" and from "s" to the end of the phrase were included.

Pitch Frequency

The fundamental-frequency measurement circuit supplied an estimate of F_0 every sample period, but some of the estimates are meaningless due to a lack of voicing. A voiced–unvoiced decision was not made. To increase the likelihood of meaningful data, only the intervals "v" to "l" and "e" to "m" were chosen for F_0 feature extraction. Two features were calculated in each interval: the mean F_0 value and its variance over those several samples. This completes the list of 405 features extracted for every utterance.

Feature evaluation was performed to eliminate those features that were relatively uninformative for speaker separation. Analysis of variance was the evaluation method employed. This method was applied by Pruzansky in a speaker identification problem involving different features [3]. Analysis of variance essentially takes the ratio (called the F-ratio) between two variances. In the numerator is the variance of the means of a particular feature for different speakers. This quantity will be large if there is considerable variation from speaker to speaker, a desirable condition. In the denominator is the average of the variances of the same feature for different speakers. That is, a variance for the feature is calculated for each speaker and these variances averaged. If this quantity is small, it indicates that on the average, the values of the feature for utterances of each speaker stay close to the mean of the feature for each speaker, another desirable condition. By forming the ratio of these quantities, a single measure of each feature's usefulness is calculated. The larger the F, the better the feature. The equation defining F exactly is

$$F = \frac{\left(\sum_{i=1}^{k} r_i - k\right)\left[\sum_{i=1}^{k}\left\{\left(\sum_{j=1}^{r_i} x_{ij}\right)^2 / r_i\right\} - \left(\sum_{i=1}^{k}\sum_{j=1}^{r_i} x_{ij}\right) \Big/ \left(\sum_{i=1}^{k} r_i\right)\right]}{(k-1)\left[\sum_{i=1}^{k}\sum_{j=1}^{r_i} x_{ij}^2 - \sum_{i=1}^{k}\left\{\left(\sum_{j=1}^{r_i} x_{ij}\right)^2 / r_i\right\}\right]}$$

where

x_{ij} feature value for jth utterance of ith speaker
r_i number of utterances of ith speaker
k number of speakers.

The quantities r_i are not equal for two reasons. First, as mentioned before, not all speakers produced the same number of utterances and some others were eliminated because of missing segmentation points. In these cases, no features existed for that utterance. Second, it is possible for some, but not all, of the features to be missing in an

TABLE IV

Testing Rank Correlation Between Pairs of Analysis of Variance Rank Orders

	Kendall (t)				Spearman (r)				
	S1U1	S2U1	S1U2	S2U2		S1U1	S2U1	S1U2	S2U2
S1U1	1.00	0.54	0.69	0.54	S1U1	1.00	0.73	0.86	0.73
S2U1	0.54	1.00	0.49	0.71	S2U1	0.73	1.00	0.66	0.88
S1U2	0.69	0.49	1.00	0.52	S1U2	0.86	0.66	1.00	0.70
S2U2	0.54	0.71	0.52	1.00	S2U2	0.73	0.88	0.70	1.00

Note: 1.00 = perfect correlation. 0.00 = random and independent. S1, S2 = speaker groups 1 and 2. U1, U2 = utterance groups 1 and 2.

otherwise acceptable utterance. For example, the algorithm for finding formants sometimes finds no peak and must indicate that the formant feature does not exist. This condition is flagged in the feature by insertion of a unique value, such as a negative formant amplitude. The feature evaluation program must take this into account and not allow the negative value to enter into mean and variance calculation. Similarly, during later training and recognition, averages were substituted for missing features in order not to bias the dot product toward either real or impostor category.

In all, four different analyses of variance were carried out. The purpose of separate tests was to check the consistency of rank from one group of utterances to another, both within the same speakers and across different groups of speakers. A base of 50 speakers was first divided in half. The utterances from each of the resulting groups of 25 speakers were further halved so that each of two groups contained about ten utterances from each speaker. Therefore, the four analyses of variance were each based on about 250 different utterances. The result of each analysis of variance was an ordered list of the 405 features based on the F-ratio.

For computational convenience, a subset of 100 of the features was extracted and the four rank orders of these compared. The comparison tool used was the Kendall and the Spearman coefficients [7]. Both coefficients are designed to produce a value of zero if the lists being compared are randomly ordered with respect to each other, and a value of unity if they are in exactly the same order. Table IV shows these results for all pairs of rankings. To calculate significance, the least significant value (Kendall $(t) = 0.49$) was converted to a normal scale and it was found that the probability of attaining a value even that large by chance was the same as the probability that a normally distributed variable will be greater than 7.2 standard deviations from its mean, an infinitesimal probability. One must conclude that all of the rank orders produced by analysis of variance are significantly correlated and that ten utterances from each of 25 speakers are sufficient to evaluate features with confidence that the results will be applicable to other utterances and other speakers.

The four rank order lists were combined in the manner suggested by Kendall, namely by assigning a composite rank for each feature based upon the average of its rank in the four original lists. This composite list was used for all later feature value judgments. For example, the top 200 of the 405 were used in most of the experiments to be reported. These are the features indicated by shading in Fig. 4(A)–(C). The obvious weakness of this and most other simple ranking techniques of feature evaluation is the lack of consideration of feature dependence. It is true that each of the shaded features is individually better for speaker verification than any of the nonshaded features by an objective measure. It is probably not true that the set of 200 shaded features is better as a whole than all other sets of 200 features chosen from the 405. Consideration of this and other related problems is given elsewhere [5].

V. Decision Experiments

The preceding sections have indicated how the utterances are digitized and segmented and how suitable features are extracted and ranked. At this stage, the decision experiments can be carried out. The results of these experiments will show the validity of the whole procedure. If verification of speakers from their voices is at all feasible by this method, it should be apparent from these results.

Once again, to make the experiments somewhat significant, many utterances from a large number of speakers should be used. Then, some confidence can be placed in the results.

In the present case, there are 50 real speakers whose utterances are used to test the validity of the proposed approach. As described in Section II, a total of approximately 100 utterances spanning a period of five recording sessions is available from each of these speakers. These utterances are divided into two equal groups of 50 for the purposes of eventual training and recognition.

1) First group of real speakers' utterances. Those collected during first two sessions and first half of the third session.

2) Second group of real speakers' utterances. Those collected during the remaining part of the third session and during the last two sessions.

Beside these 50 speakers, some utterances from a separate set of 68 speakers are also available. These are the impostors. Two groups are again formed.

1) First group of impostor utterances. About nine utterances from each of 29 speakers assigned to the impostor category.

2) Second group of impostor utterances. All utterances (about 20 from each) from the remaining 39 speakers assigned to the impostor category.

A modified version [1] of the classical Adaline technique is utilized for the training and recognition procedure inherent in the decision process. During training, this technique employs several training utterances and creates a reference profile or weight vector W for each of the real speakers. The training utterance set consists of the first group of utterances from the real speaker for which the weight vector is desired and the first group of impostor utterances. The recognition set is comprised of the second group of utterances from the same real speaker and the second group of impostor utterances. After the weight vector is formed for a real speaker, recognition is performed by using this vector and the recognition set of utterances. Thus, the recognition data is completely separate from the training data.

The weight vector W is such that, for a convergent training procedure [1],

$$W \cdot R > K \mid W \mid$$

and

$$W \cdot I < -K \mid W \mid$$

where R is any utterance from the real training set, I is any one from the impostor training set, K is some finite constant, and the symbol $(\cdot)$ implies inner product operation. The value of K was set to 5 for the current experiments. This choice was a compromise and was made after some preliminary investigations on convergence time, generalization ability, and other such related factors [1].

During recognition, the inner products of the weight vector with the recognition utterances are determined. A threshold is then selected so that the percentage of misclassified real utterances is equal to the percentage of misclassified impostor utterances. This single percentage is plotted in Fig. 5 in histogram form. It is clear that except for a small number of speakers, the error rate of misclassification is generally low. Mean error rate is one percent. An investigation carried out on the speaker with the highest error rate has revealed that this high error rate can be attributed largely to segmentation inaccuracies. It is possible that an improved segmentation strategy will decrease the overall error rate on all speakers. In this connection, it may be mentioned that the effect of segmentation error is serious for the utterances from the real speaker only. Such utterances with wrong segmentation points lead to wrong feature sets and are likely to produce wrong decisions. On the other hand, erroneous feature sets for impostor utterances resulting from inaccurate segmentation are not likely to cause these utterances to be misclassified as utterances from the real speakers.

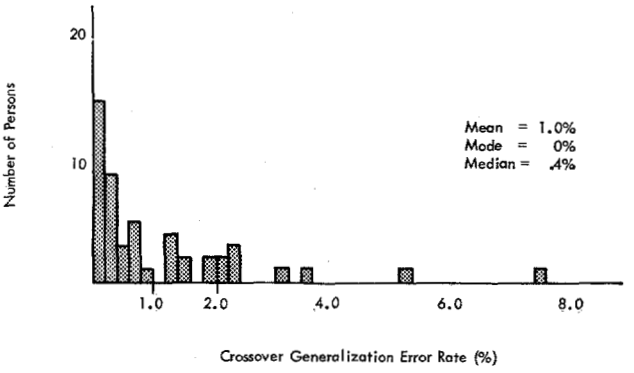


Fig. 5. Generalization error histogram.

Another experiment using six real speakers will now be reported. The objective is to test the hypothesis that utterances collected during one session only are less desirable for training purposes than the utterances obtained during a larger number of sessions. The following rationale was employed for forming such a hypothesis. One may argue that the utterances spoken during one session tend to be rather "similar" to one another and do not depict typical variations of speech patterns that may be expected at later instances of time. Thus, the weight vector derived from the utterances of one session is not likely to be representative. However, if utterances are collected in at least two sessions which are separated by a prudent time interval, many of the person's possible speech variations are encountered and the weight vector is likely to be more representative. No investigations about what constitutes a suitable time interval between two sessions have been made in the following experiment.

In the first part of the experiment, only the first 20 utterances produced by each real speaker in the first session are used for training. Training utterances of the impostors and recognition utterances of both the real and the impostors are unchanged from the earlier experiments. In previous terminology, only the first group of utterances from the real speaker is different; the second group of real utterances and the two groups of impostor utterances are the same as before. The error rates for the six real speakers are shown in the row marked "1–20" of Table V. The column marked "Average" gives the average errors rates across the six speakers.

TABLE V

Results with Various Number of Training Utterances

Real Utterances of Group I	Speaker 1 Percent	Speaker 2 Percent	Speaker 3 Percent	Speaker 4 Percent	Speaker 5 Percent	Speaker 6 Percent	Average Percent
1–20	9.1	2.0	1.7	1.6	5.9	0	3.4
1, 3, · · · , 39	3.6	1.4	0.5	0.5	3.5	0	1.6
1–50	3.1	1.2	2.1	0.8	3.4	0	1.8

In the second part of the experiment, only 20 utterances of the real speaker are again used for training, but, this time, they are the alternate 20 utterances from the first *two* sessions, numbered 1, 3, 5, · · · , 39. The total number of training utterances is kept constant so that it does not enter as a variable in the experiment. Other three groups of utterances are held fixed as before. The error rates are shown in the row labeled "1, 3, · · · , 39" of Table V. The number in the last row of this table marked "1–50" are the original error rates for these six speakers when the first 50 utterances from each of them are used for training. Comparing the three rows of results, it may be observed that the utterances of one session represented in the row "1–20" indeed lead to higher error rates than the utterances obtained during two sessions shown in the row "1, 3, · · · , 39." Utterances of *three* sessions which results in the error rates of the row "1–50" seem to have no added value over the utterances of two sessions. Occasional discrepancies in the general trend of results were noted, but no explanations for these occurrences were apparent.

Finally, it is pointed out that these conclusions about the number of sessions suitable for gathering training utterances may depend on the particular speakers employed, instructions given to them, and other such factors as mentioned in Section II.

VI. Comments and Further Work

The experiments reported in this paper and in [1] have used a realistic data base which is believed to be larger than any other data bases previously reported in the same area of work. This data base permits the experimenter to draw more meaningful conclusions and grants him considerable freedom in satisfactory implementations of various algorithms.

One problem pointed out in Section III is the relatively high (ten percent) phrase-rejection rate. This rate could possibly be lowered in at least two ways. First, it is necessary to invent more effective segmentation techniques. This means that events in more utterances will be successfully detected leading to a lower rejection rate. Next, a technique for dealing with utterances which has one or more missing events needs to be found. This will also yield reduced phrase-rejection rate.

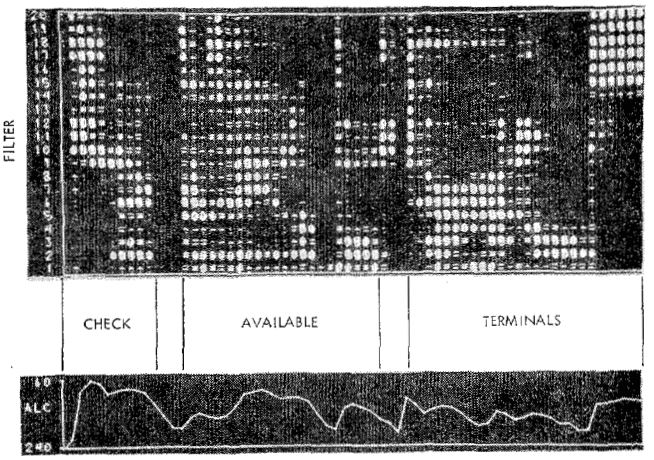


Fig. 6. Spectrogram display in IBM 2250.

A possible way of improving the segmentation procedure is by adopting a strategy of detecting individual events independently. Later, these events may be correlated to verify that they occur in their expected sequence.

Considerable facility in the design of a segmentation strategy may be gained by utilizing an interactive display system. The system should have the provision for displaying the results obtained from the experimenter's program in a convenient form as well as the digitized spectrogram of the speech. Then, the user could modify various parameters of his program on-line and observe the effect in a short time. A good understanding of the capabilities and limitations of his program may be obtained in this manner. Fig. 6 shows a digital spectrogram display in an IBM 2250.

One of the important aspects of speaker verification missing in this research is the utilization of the utterances of female speakers. It is quite likely that the segmentation and feature selection strategies developed in the present paper will be inadequate for the female speech. Thus, a study of female utterances should be undertaken. Other important areas of research in the verification problem may be mentioned. Investigations regarding the effect of colds or other symptoms that may affect speech quality can be carried out. Sensitivity of the verification system

to impostors trying intentionally to mimic a real speaker needs to be studied. Such impostors should include speakers with formal training in speech since they are effectively able to control various characteristics of their voice. Siblings of real speakers should also be considered since their voice may be "similar" to the voice of the real speakers due to their possibly similar physical speech apparatus or environmental background.

In summary, effective automatic speaker verification has been demonstrated, the results of collecting training utterances over one or more number of sessions have been investigated, and some further areas of research have been pointed out.

Acknowledgment

The authors wish to thank their associates in the Speech Processing and other Advanced Technology Departments, for invaluable assistance in theory, hardware design and construction, data gathering and preprocessing, and programming. Special credit is due Dr. N. R. Dixon who composed the five phrases listed in Table I, designed the data-taking paradigm, advised the authors regarding many aspects of the feature extraction process, and helped check the acceptability of the segmentation results.

References

[1] S. K. Das and W. S. Mohn, "Pattern recognition in speaker verification," in *1969 Fall Joint Computer Conf., AFIPS Conf. Proc.*, vol. 35, pp. 721–732.

[2] K. P. Li, J. E. Dammann, and W. D. Chapman, "Experimental studies in speaker verification, using an adaptive system," *J. Acoust. Soc. Am.*, vol. 40, Nov. 1966, pp. 966–978.

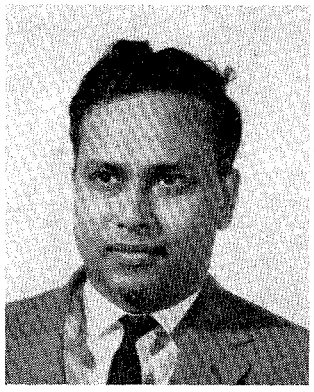
[3] S. Pruzansky, "Talker-recognition procedure based on analysis of variance," *J. Acoust. Soc. Am.*, vol. 36, Nov. 1964, pp. 2041–2047.

[4] J. E. Luck, "Automatic speaker verification, using cepstral measurement," *J. Acoust. Soc. Am.*, vol. 46, Oct. 1969, pp. 1026–1032.

[5] W. S. Mohn, "Statistical feature evaluation in speaker identification," Ph.D. dissertation, Dept. Elec. Eng., North Carolina State Univ., Raleigh, July 1969.

[6] J. W. Glenn and N. Kleiner, "Speaker identification based on nasal phonation," *J. Acoust. Soc. Am.*, vol. 43, Feb. 1968, pp. 368–372.

[7] M. G. Kendall, *Rank Correlation Methods.* New York: Hafner, 1962.



Subrata K. Das was born in Contai, West Bengal, India, on April 18, 1940. He received the B.E.E. degree from Jadavpur University, Calcutta, India, in 1960, the M. Tech. degree from the Indian Institute of Technology, Kharagpur, India, in 1961, and the Ph.D. degree in electrical engineering from the University of Arizona, Tucson, in 1966.

In 1962 he was with Philips India Ltd., Calcutta, India, developing electronic measuring instruments. During the summers of 1964 and 1965 he was with Argonne National Laboratory, Argonne, Ill., working on the electrical feedback system of the Zero Gradient Synchrotron. Presently he is a Staff Engineer with IBM Corporation, Raleigh, N. C., where he works on speaker verification and other pattern recognition projects. He has also done work in the areas of speech processing techniques, fast Fourier transforms, and adaptive processes.

William S. Mohn was born in St. Louis, Mo., on July 1, 1941. He received the B.S. and M.S. degrees in electrical engineering simultaneously from the Massachusetts Institute of Technology, Cambridge, in 1964, and the Ph.D. degree in electrical engineering from North Carolina State University, Raleigh, in 1969.

While at M.I.T. he was enrolled in the cooperative program with concurrent work assignments at IBM in the areas of circuit design, optical character recognition, and time-domain telephone line equalization. Since 1964 he has been employed by the IBM Corporation, Kingston, N. Y. and Research Triangle Park, N. C. His primary interests have been the development of techniques for speaker verification, including feature extraction, feature evaluation, and decision techniques.

Dr. Mohn is a member of Eta Kappa Nu, Tau Beta Pi, and Sigma Xi.