AD NUMBER

ADB008708

LIMITATION CHANGES

TO:

Approved for public release; distribution is unlimited.

FROM:

Distribution authorized to U.S. Gov't. agencies only; Test and Evaluation; OCT 1975. Other requests shall be referred to Rome Air Development Center, Griffiss, AFB, NY 13441.

AUTHORITY

RADC ltr, 12 Apr 1978

RADC-TR-75-264
Final Technical Report
October 1975

AUTOMATIC CLASSIFICATION OF LANGUAGES

Texas Instruments Incorporated

Rome Air Development Center
Air Force Systems Command
Griffiss Air Force Base, New York 13441

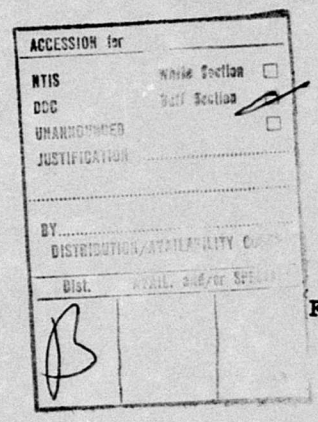
This report has been reviewed and approved for publication.

APPROVED: Richard S. Vonusa

RICHARD S. VONUSA
Project Engineer

APPROVED: H Davis

HOWARD DAVIS
Technical Director
Intelligence & Reconnaissance Division



FOR THE COMMANDER: John P. Huss

JOHN P. HUSS
Acting Chief, Plans Office

Do not return this copy. Retain or destroy.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Date Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: RADC-TR-75-264	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle): AUTOMATIC CLASSIFICATION OF LANGUAGES	5. TYPE OF REPORT & PERIOD COVERED: Final Technical Report, May 74 - May 75
6. PERFORMING ORG. REPORT NUMBER: N/A	
7. AUTHOR(s): R. Gary Leonard; George R. Doddington	8. CONTRACT OR GRANT NUMBER(s): F30602-74-C-0245
9. PERFORMING ORGANIZATION NAME AND ADDRESS: Texas Instruments Incorporated, 13500 North Central Expressway, Dallas TX 75222	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 31011F; 70550712
11. CONTROLLING OFFICE NAME AND ADDRESS: Rome Air Development Center (IRAP), Griffiss AFB NY 13441	12. REPORT DATE: October 1975
	13. NUMBER OF PAGES: 39
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office): Same	15. SECURITY CLASS. (of this report): UNCLASSIFIED
16. AF-7055	17. 705507
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE: N/A

16. DISTRIBUTION STATEMENT (of this Report)

Distribution limited to U. S. Gov't agencies only; test and evaluation; October 1975. Other requests for this document must be referred to RADC (IRAP), Griffiss AFB NY 13441.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

Same

18. SUPPLEMENTARY NOTES

RADC Project Engineer:
Richard Vonusa (IRAP)

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Language Classification
Speech
Pattern Recognition

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Studies were made of language classification algorithms which were based on the recurrence frequencies of sequences of phoneme-like sound segments. Sequences of length 2, 3, 4 and 5 were considered. The sequence recurrences comprised acceptances of acoustic hypotheses, which in turn are based on a time-frequency scanning error measure, and on occurrence time relationships. Classification performance was estimated using 50 independent test speakers of five languages. Individual language accuracies ranged from 17 to 90 percent, with the overall five-language accuracy being 70 percent.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Date Entered)

PREFACE

This Final Report describes research on automatic language classification by Texas Instruments Incorporated, Equipment Group, 13500 North Central Expressway, Dallas, Texas, under Contract No. F30602-74-C-0245 for Rome Air Development Center, Griffiss Air Force Base, New York. Mr. Richard S. Vonusa (IRAP) was the RADC Project Engineer. The report covers work performed from May 1974 through May 1975.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

SECTION I

INTRODUCTION

The problem studied is to design and simulate a system that will automatically determine to which of several specified languages a given segment of speech belongs, and to do this with small probability of error. This report contains the results of the second phase of study of such a system. In the first phase [Reference 1] classification was based on language likelihoods computed for certain reference sounds. Those sounds were short phoneme-like segments. In the second phase, classification is based on sequences of several such segments, which allows more accuracy and reliability in the automatic segmentation process. Another improvement is the use of "time-frequency scanning" to accept or reject hypothesized occurrences of component sound segments.

The algorithms discussed here automatically produce the information needed for language discrimination without reference to the particular languages being considered. This approach allows treatment of additional languages with little additional effort and with no need for special knowledge of those languages.

Two separate data sets of equal size have been used in this study. The first is used to design the classifier, and the second is used for estimating the probability of error to be expected in using the classifier. The data sets used in this second phase of the study are the same as those used in the first phase.

A decision rule based on occurrence frequency information about sequences of length 5 and single segments allowed five-language classification accuracy of 70 percent.

SECTION II

DATA PREPROCESSING

Analog speech data recordings for this study were provided by RADC. The data are from five languages, denoted L_1, L_2, L_3, L_4, and L_5. This designation is adequate since the algorithms and techniques described herein treat each language identically. No processing is tailored to specific languages. Data from 100 distinct speakers were processed: 50 speakers provided data for estimating decision parameters and generating reference files ("training data"); and data from 50 other speakers ("testing data") were used to provide unbiased estimates of decision accuracy.

The training data consisted of 90-second segments of speech from each of 10 speakers of each of the five languages. The testing data comprise 90-second segments from: 10 speakers of L_1, L_3, and L_5; 6 speakers of L_2; and 14 speakers of L_4. The testing data were used only to determine performance.

1. ANALOG PREPROCESSING

The analog speech data base was preprocessed using the hardware shown functionally in Figure 1. Fifteen bandpass filters were used to provide a time-frequency signal analysis. The center frequencies and bandwidths of these filters are shown in Table I. Following the low-pass filtering, the signals in each channel are sampled, digitized to 11 bits, and stored for additional processing. One hundred samples per second are retained to represent the speech information. In the following description, g will denote a column vector of data values stored at some specified time. Since the operations performed at each sampling time are identical, reference to the specific time will be suppressed. The symbol $g' = [g_1\ g_2 \ldots g_k]$ will denote the transpose of the column vector g.

2. NORMALIZATION AND QUANTIZATION

Some speaker normalization is accomplished by regressing the data vector g upon regression vectors chosen to maximize between-speaker to within-speaker variance in the time-frequency spectrum [Reference 1]. The expression for the normalized data vector is

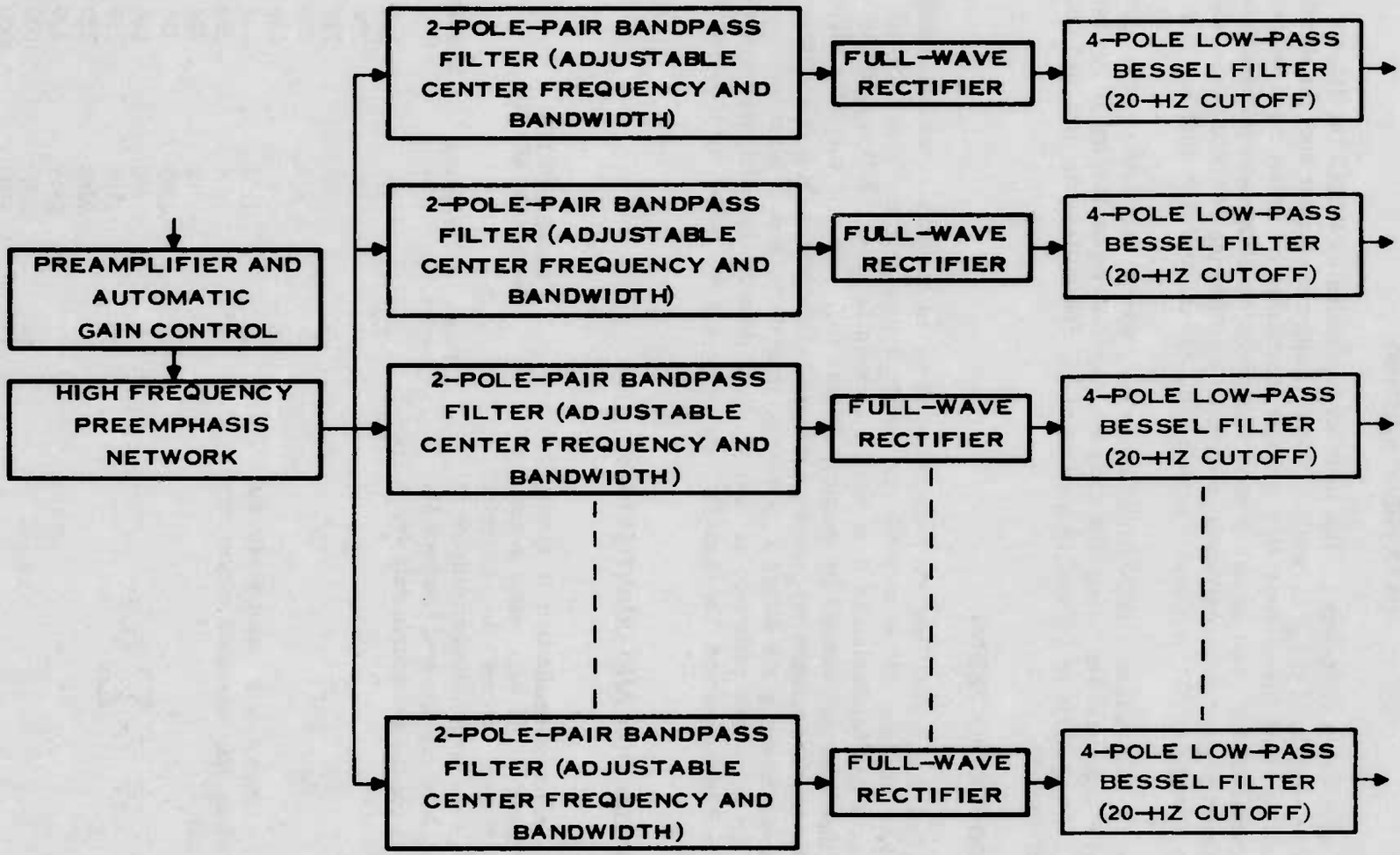
$$g_N = 1/\sigma\ \tilde{f}$$

where $\tilde{f}$ is the original data vector g with the components along the regression vectors subtracted out, and

$$\sigma = \tilde{f}' \cdot \tilde{f} = \sum \tilde{f}_i^2$$

TABLE I. FILTER BANK PARAMETERS

Filter Number	Frequency (Hz)	Bandwidth (Hz)
1	355	220
2	530	220
3	705	220
4	880	220
5	1055	220
6	1230	220
7	1405	220
8	1580	220
9	1755	220
10	1930	220
11	2105	220
12	2280	220
13	2455	220
14	3500	1000
15	6500	3000



174845A

Figure 1. Functional Block Diagram, Analog Spectral Preprocessor

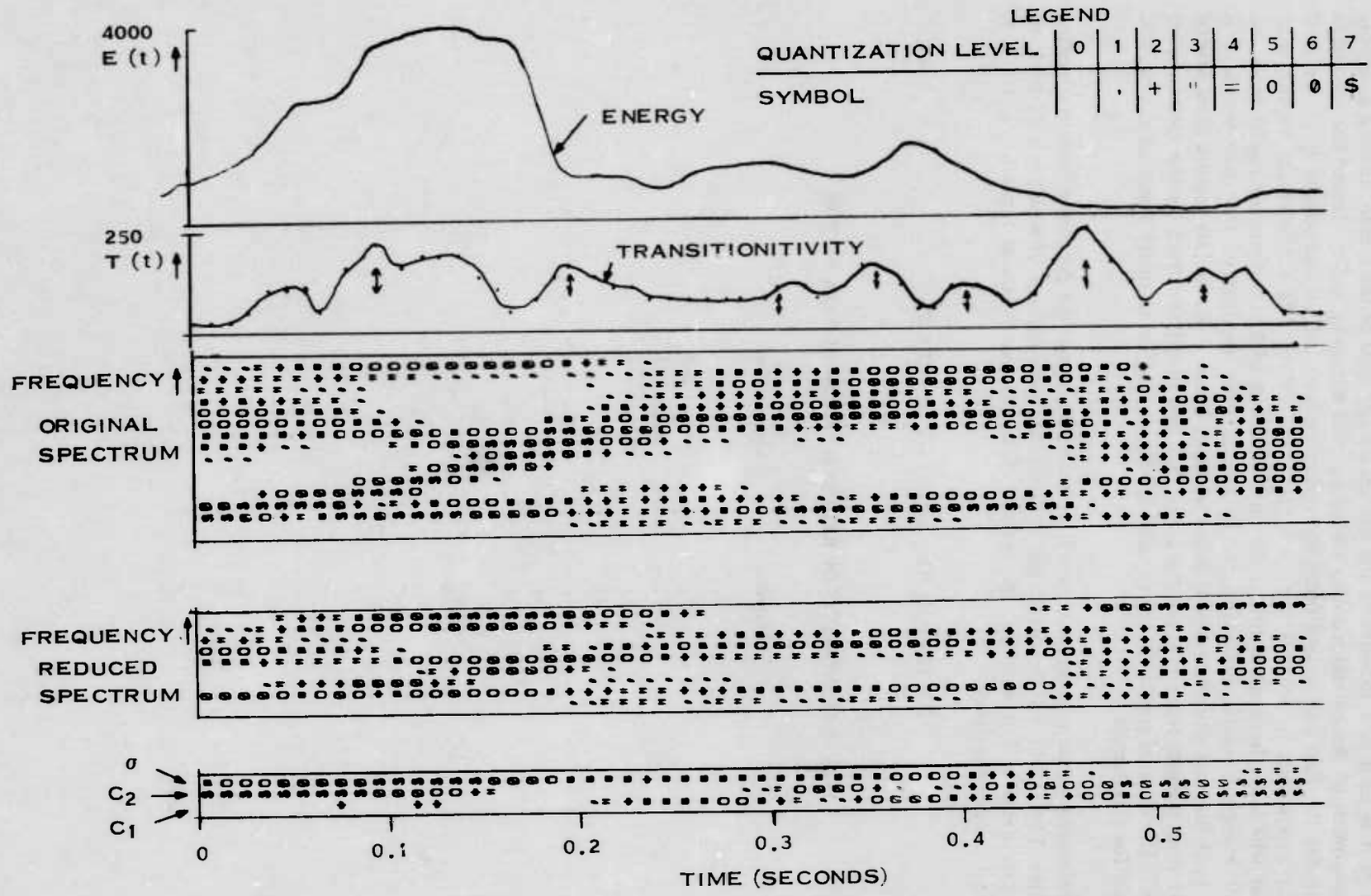
The value of σ is used as a measure of the overall energy level of the speech. Letting g_i denote the *ith* component of the normalized data vector g_N, and letting c_1 and c_2 denote the regression coefficients for the first and second order regression vectors, the *jth* component f_j of the final data vector f is shown in Table II, for $j = 1, 2, \ldots, 12$. Thus, f is a 12-element vector, nine elements of which are from normalized filter outputs, along with two regression coefficients and the overall energy measure. Following normalization, each element of the data vector f is quantized to 3 bits and stored on digital magnetic tape. Figure 2 shows data vectors from several consecutive time samples forming a quantized and digitized speech spectrum for the spoken word "warheads." The original spectrum, the reduced spectrum, and the auxiliary measures c_1, c_2, and σ are shown for this sample.

In following sections, computations will be made which use the data from certain specified time samples. These data will be considered as a matrix, each column of which is a 12-tuple of the form just defined. For example, the matrix of preprocessed data at times $t - 2$, $t - 1$, t, $t + 1$, and $t + 2$ will be written

$$P = [f(t-2)\ f(t-1)\ f(t)\ f(t+1)\ f(t+2)]$$

TABLE II. COMPONENTS OF REDUCED SPECTRUM DATA VECTOR

Component	Composition
f_1	g_1
f_2	$\frac{1}{2}(g_2 + g_3)$
f_3	$\frac{1}{2}(g_4 + g_5)$
f_4	$\frac{1}{2}(g_6 + g_7)$
f_5	$\frac{1}{2}(g_8 + g_9)$
f_6	$\frac{1}{2}(g_{10} + g_{11})$
f_7	$\frac{1}{2}(g_{12} + g_{13})$
f_8	g_{14}
f_9	g_{15}
f_{10}	c_1
f_{11}	c_2
f_{12}	σ



191421

Figure 2. Spectral Representation of Speed Data From the Word "Warheads," and Auxiliary Measures

SECTION III

KEY SOUND DETECTION

The general solution of the language classification problem is to first find those acoustic elements which have language characterizing capability (i.e., those elements which occur with greatly different likelihoods in different languages), and then, having observed such an acoustic element, estimate the likelihoods of the various hypothesized languages and choose that language which is most likely. This section deals with the first step of this solution, the isolation of key sounds. The structure of the sounds considered in this study is a sequence of phoneme-like acoustic elements, as determined from the spectral representation of the speech data described in the previous section. Each sequence of length k is determined by k points of time registration, and is represented by k sets of data vectors, where k = 2, 3, 4, or 5. The result of the procedures described in this section is a reference set of key sounds (sequences) to be used for estimating language likelihoods.

1. INTERSEGMENT SIMILARITY

Extensive use is made of the squared error between two matrices (vectors) representing sound segments. Suppose $F = [f_i(j)]$ and $G = [g_i(j)]$ each comprise M data vectors, $i = 1, 2, \ldots, 12$; $j = 1, 2, \ldots, M$. Then the squared error between F and G, written e(F,G), is defined to be

$$e(F,G) = \sum_{j=1}^{M} \sum_{i=1}^{12} [f_i(j) - g_i(j)]^2$$

2. TIME REGISTRATION

Points of time registration in the data are defined in terms of the overall energy measure, σ, and a transitionitivity function, T, a real-valued function of time which reflects the magnitude of dynamic spectral change. Let R(t) denote the 12 X3 matrix consisting of data vectors from three consecutive sampling times centered at time t; i.e.,

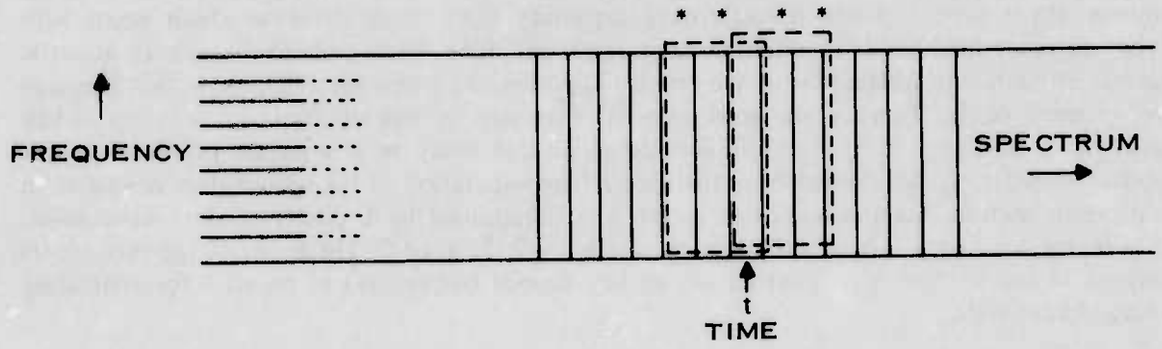
$$R(t) = [f(t-1)\ f(t)\ f(t+1)]$$

Then, the transitionitivity at time t, T(t), is defined to be the squared error between R(t − 1) and R(t + 1):

$$T(t) = e[R(t-1), R(t+1)]$$

Figure 3 illustrates the format of the data vectors used to compute T. If T(t) is small, then the two matrices are similar and t is a time of relatively steady-state speech. If T(t) is large, then t is a time of transition in the spectrum.

During data processing, the T function is computed at each sample time and is monitored to determine its peaks and valleys. Each peak in the T function is labeled as a time registration point, provided that (1) the value of the peak is greater than 50, and (2) the smallest values of the overall energy measure σ in a 0.1-second neighborhood about that time is greater than 280. Condition (1) is imposed to overlook those small spectral changes which probably are not actual



191422

Figure 3. Data Vectors Used to Compute Transitionitivity

phoneme boundaries, and condition (2) prevents consideration of silence segments. The threshold levels were determined as a result of inspection of speech spectra.

At each registration time, a scanning pattern (SP) is extracted and stored for use in representing candidate reference sequences. Let S(t) denote a typical SP extracted at time t. Then S(t) consists of three derived data vectors

$$S(t) = [g_1 \; g_2 \; g_3]$$

where

$$g_1 = \tfrac{1}{2}[f(t-2) + f(t-1)]$$

$$g_2 = \tfrac{1}{2}[f(t+1) + f(t+2)]$$

$$g_3 = g_1 - g_2 .$$

The data vectors marked with asterisks in Figure 3 are used in determining an SP at time t.

The first 90 seconds of speech data from each of the 50 training speakers was processed to extract and store scanning patterns at each time registration point. Table III shows the numbers of SPs extracted.

TABLE III. NUMBERS OF EXTRACTED SCANNING PATTERNS

Language	L_1	L_2	L_3	L_4	L_5
Number of Scanning Patterns	5,399	6,759	5,012	5,699	6,345

3. REFERENCE FILE GENERATION

Any sequence of k(k = 2, 3, 4, and 5) consecutive scanning patterns separated by no more than 0.25 second was considered to represent a potentially useful sound sequence. The total number of these candidate reference sequences (approximately 116,000, as seen from Table III) was too large to process the desired amount of training and testing. Hence, the procedures in the following sections were adopted to choose a small subset of these candidate reference sequences for use in the final language classifier. The object of the procedures is to find those sequences which occur often in the data and are distinct from each other. The procedures to be described were performed for each value k, k = 2, 3, 4, and 5. Hence, specific reference to the particular sequence length will often be omitted.

a. Scanning Error

In ascertaining the occurrence of a sequence, use is made of the "scanning error," E, a real-valued function of a fixed scanning pattern, S, and of time, t. Let F(t) denote the derived data matrix $[g_1\ g_2\ g_3]$, where

$$g_1 = \tfrac{1}{2}[f(t-2) + f(t-1)]$$

$$g_2 = \tfrac{1}{2}[f(t+1) + f(t+2)]$$

$$g_3 = g_1 - g_2$$

$f(t)$ = data vector at time t.

Note that F has the same format as the scanning pattern. Then the scanning error E(S,t) for scanning pattern S is defined to be

$$E(S,t) = e[S, F(t)]$$

A relative minimum in the scanning error for S indicates a time at which the speech data is similar to the scanning pattern S.

b. Sequence Detection

The recurrence of reference sequences is detected by first scanning the input data to hypothesize the occurrence of appropriate time registration points, and then hypothesizing sequence recurrence when the relationships among hypothesized time registration points correspond to those in the reference sequence. Rejection of sequence recurrence is based on spectral similarity between reference sequence scanning patterns and the input data.

To be more specific, consider the detection of reference sequences of length 2. Let (S_1, S_2) denote a reference pair, where S_i is the scanning pattern extracted from the training data at time T_i, i = 1, 2. Let $\Delta\hat{t}$ denote $|T_2 - T_1|$, the time separation between scanning patterns, and

assume that $T_1 < T_2$. During processing of the input data, the scanning errors $E(S_i, t)$ $i = 1, 2$ are computed and monitored to label the valleys in each of these scanning error functions. These valley points are hypothesized to be time registration points for the corresponding scanning pattern. Whenever a valley in $E(S_1, t)$ precedes a valley in $E(S_2, t)$ by less than 0.25 second, the recurrence of the reference pair (S_1, S_2) is hypothesized.

The basis for acceptance or rejection of this hypothesis is the pair error, denoted E_2. Let t_i denote the time of occurrence of a valley in $E(S_i, t)$, $i = 1,2$, and let $\Delta t = |t_2 - t_1|$. Let $e_i = E(S_i, t_i)$ denote the value of the scanning error at the valley time t_i, $i = 1,2$. Assume that $|t_1 - t_2| < 0.25$ second, so that the recurrence of (S_1, S_2) is hypothesized. This hypothesis is rejected if and only if $E_2 > EMAX_2$, where $EMAX_2$ is a fixed threshold and E_2 is defined to be

$$E_2 = \frac{(e_1 + 40)(e_2 + 40)}{2048} \left\{ 1 + \frac{4|\Delta t - \Delta \hat{t}|}{\max(5, \Delta \hat{t})} \right\}$$

It can be seen that, for detecting the occurrence of a reference pair (S_1, S_2), the time separation of registration times in the data must be close to that of the scanning patterns, and also the scanning patterns must each be similar to corresponding transitions found in the data.

For detection of sequences of length k, k = 3, 4, and 5, the scanning error valleys for each of the k scanning patterns are recorded, and an occurrence hypothesized whenever (1) k scanning error valleys, one from each scanning pattern, occur in the same order as did the scanning patterns; and (2) no time interval between adjacent valleys is greater than 0.25 second. The pair errors between each of the $k - 1$ pairs of valleys are summed to form a k-sequence error, E_k. The hypothesized occurrence of the sequence is rejected if and only if $E_k > EMAX_k$, where $EMAX_k$ is a specified threshold for sequences of length k.

c. Preliminary Sequence Selection

Because the average processing time needed to detect occurrences of a single reference sequence in all the training and testing data is approximately 20 minutes, most of the approximately 116,000 candidate reference sequences must be eliminated from consideration. As a first step toward this goal, a study was made to determine what constitutes similarity among candidate reference sequences. Let $P_1 = (S_{11}, S_{12})$ and $P_2 = (S_{21}, S_{22})$ denote candidate reference sequences, where S_{ij} is a scanning pattern which occurred at time t_{ij}, $i, j = 1,2$. Let e_1 and e_2 denote the squared error $e(S_{11}, S_{21})$ and $e(S_{12}, S_{22})$, respectively, and define $\Delta t = |t_{12} - t_{11}|$, and $\Delta \hat{t} = |t_{22} - t_{21}|$. Then define the similarity between P_1 and P_2 to be

$$E_s(P_1, P_2) = \frac{(e_1 + 40)(e_2 + 40)}{2048} \left\{ 1 + \frac{4|\Delta t - \Delta \hat{t}|}{\max(5, \Delta \hat{t})} \right\}$$

Note the intentional analogy to the pair error E_2.

Let $T_1 = \{S_i: i = 1, 2 \ldots, 100\}$ denote a set containing the first 100 candidate reference sequences of length k hypothesized from the first training speaker of language L_2. A cumulative frequency distribution was plotted for the values of the similarities $E_s(S_i, S_j)$, $1 \leq i < j \leq 100$. A similarity level β was determined such that $P\ E_s(P_i, P_j) < \beta = 0.75$. The values of β determined for each k are shown in Table IV.

TABLE IV. SIMILARITY THRESHOLD

Sequence Length	2	3	4	5
Threshold	49	92	132	176

Then, all the candidate reference sequences from each language were ordered according to the following procedure. Let the language and the sequence length be fixed. A subset T_2 consisting of 10 percent of the candidate reference sequences from each of the 10 training speakers was formed. For every candidate reference sequence S from the language considered, the similarities $E_s(S, R)$, $R \epsilon T_2$, were computed, and the number N(R) of similarity values less than β (for sequence R) were noted. All candidate reference sequences were then ordered according to N, with the first sequence being the one with the highest value of N. This procedure places that sequence occurring more frequently higher in the ordering.

A reduced file of candidate reference sequences is formed for each language and each sequence length by first placing in the file that sequence which is first in the above ordering, and then adding succeeding sequences in the ordering, provided that the minimum of the similarities between the sequence to be added and each sequence already in the reduced file is greater than a fixed threshold. This ensures that each sequence in the reduced file is distinct from all others in the file.

The procedure for determining the add threshold is as follows. A small file of candidate reference sequences from 34 consecutive scanning patterns from one training speaker was formed. Occurrences of these sequences were detected in 10 seconds of data from each of the nine other training speakers of the language. The sequence rejection threshold $EMAX_k$ was set large enough so that all hypothesized sequences and their corresponding E_k values are recorded. The cumulative frequency distribution of the values of E_k was plotted and the add threshold α was chosen such that $P\{E_k < \alpha\} = 0.98$. Table V shows the add thresholds for each language and each sequence length.

The total numbers of hypothesized sequences from the experiment described in the preceding paragraph are shown in Table VI for each language and each sequence length. These numbers were used to determine the proportions of the desired numbers of reference sequences of the various lengths. These proportions are the inverses of the ratios existing among the total number of hypothesized sequences. That is, if there were twice as many pairs hypothesized as there were triples, it is desired to have half as many reference pairs as reference triples in the reduced file. The total numbers of reference sequences added to the reduced file was restricted by available processing time and computer core storage limitations. The numbers of reference sequences retained in the reduced files are shown in Table VI for each language and each sequence length. The total numbers of reference sequences of each length are shown in Table VII. A total of 452 reference sequences remained in the reduced file, denoted F.

TABLE V. ADD THRESHOLDS

	Sequence Length			
Language	2	3	4	5
L_1	94	152	204	256
L_2	96	156	210	264
L_3	104	168	228	296
L_4	84	140	186	224
L_5	92	148	214	248

d. Data Processing for Occurrence Information

Both the training data and the testing data (90 seconds of speech from each of 100

TABLE VI. FILE REDUCTION PARAMETERS

Language Index	Sequence Length	Number of Hypothesized Sequences	Number of Sequences Retained
1	2	10,817	10
1	3	5,384	21
1	4	2,810	42
1	5	1,447	19
2	2	14,846	10
2	3	7,625	21
2	4	4,026	40
2	5	2,164	18
3	2	13,574	7
3	3	5,678	15
3	4	1,646	45
3	5	532	21
4	2	18,880	14
4	3	11,833	23
4	4	7,260	37
4	5	4,027	19
5	2	12,243	9
5	3	5,715	20
5	4	2,670	41
5	5	1,249	19

TABLE VII. NUMBER OF REFERENCE SEQUENCES IN REDUCED FILES

Sequence Length	2	3	4	5	Total
Number of Sequences	51	100	205	96	452

speakers) were processed to detect the occurrences of each sequence in the reference file F. The values of $EMAX_k$ (the rejection level for hypothesized sequences) was set large enough that all hypothesized sequences were accepted. For each accepted sequence, record was made of (1) the index S of the speaker whose data was being processed (and, hence, of his language L), (2) the index R of the accepted reference sequence (and, hence, its length k), and (3) the value E_k of the overall sequence error. This processing required approximately 150 hours of computer time, using a TI 980A minicomputer.

To determine the effects of varying the rejection level for detecting sequence recurrence, six sets of thresholds $EMAX_k$ were used in turn to determine an array N(R, S, L), where an entry in this array is the number of occurrences of reference sequence R during processing of data from speaker S of language L. (As previously described, an occurrence is counted whenever $E_T \leqslant EMAX_k$.) One set of thresholds (Case 0) is the one mentioned in the previous paragraph,

which yields *all* hypothesized sequences. The other sets contain successively lower thresholds, thereby requiring successively better match between reference and data to yield an occurrence. Case i, i = 1, 2, . . . , 5, are obtained as follows. First, an empirical probability density plot was obtained for values of E_T from training data for Case 0. Figures 4, 5, 6, and 7 show this density for sequence length 2, 3, 4, and 5, respectively. Let N_T denote the total number of sequences hypothesized (for some sequence length). From the corresponding distribution, threshold values were determined which would yield $N_T/2^i$ detected sequences, for Case i, i = 1, 2, 3, 4, 5. This procedure was followed for each value of sequence length. The resulting thresholds are shown in Table VIII. Analysis of the results of using the six cases are presented in Section V.

TABLE VIII. SEQUENCE ACCEPTANCE THRESHOLDS

	Case					
Length	**0**	**1**	**2**	**3**	**4**	**5**
2	*999	29	18	13	9	7
3	999	61	41	31	24	19
4	999	96	69	53	42	34
5	999	119	87	69	56	46

*Threshold of 999 allows acceptance of every hypothesized sequence.

e. Final Reference File Formation

The final reference file, F*, of sequences to be used in computing decision functions is formed by deleting from F those reference sequences which had too little language specificity. Such sequences were determined by considering the average information remaining (uncertainty, entropy) after detection of a reference sequence in the training data. The lower this uncertainty, the better is the language discrimination capability of that sequence. Specifically, the entropy H(R) associated with the detection of sequence R is

$$H(R) = -\sum_{i=1}^{5} p(L_i|R) \log [p(L_i|R)]$$

where $p(L_i|R)$ is the language likelihood, given that R has occurred. Let the symbol $\underline{S}_i$ denote the collection of training speakers of language L_i, i = 1, 2, . . . 5, and let M(R, L) denote the number of detections of reference sequence R during speech from all training speakers of language L. Then

$$M(R, L_i) = \sum_{SES_i} N(R, S, L_i), \; i = 1, 2, \ldots 5$$

The likelihood $p(L_i|R)$ is computed as

$$P(L_i|R) = \frac{q(R, L_i)}{\sum_{i=1}^{5} q(R, L_i)}$$

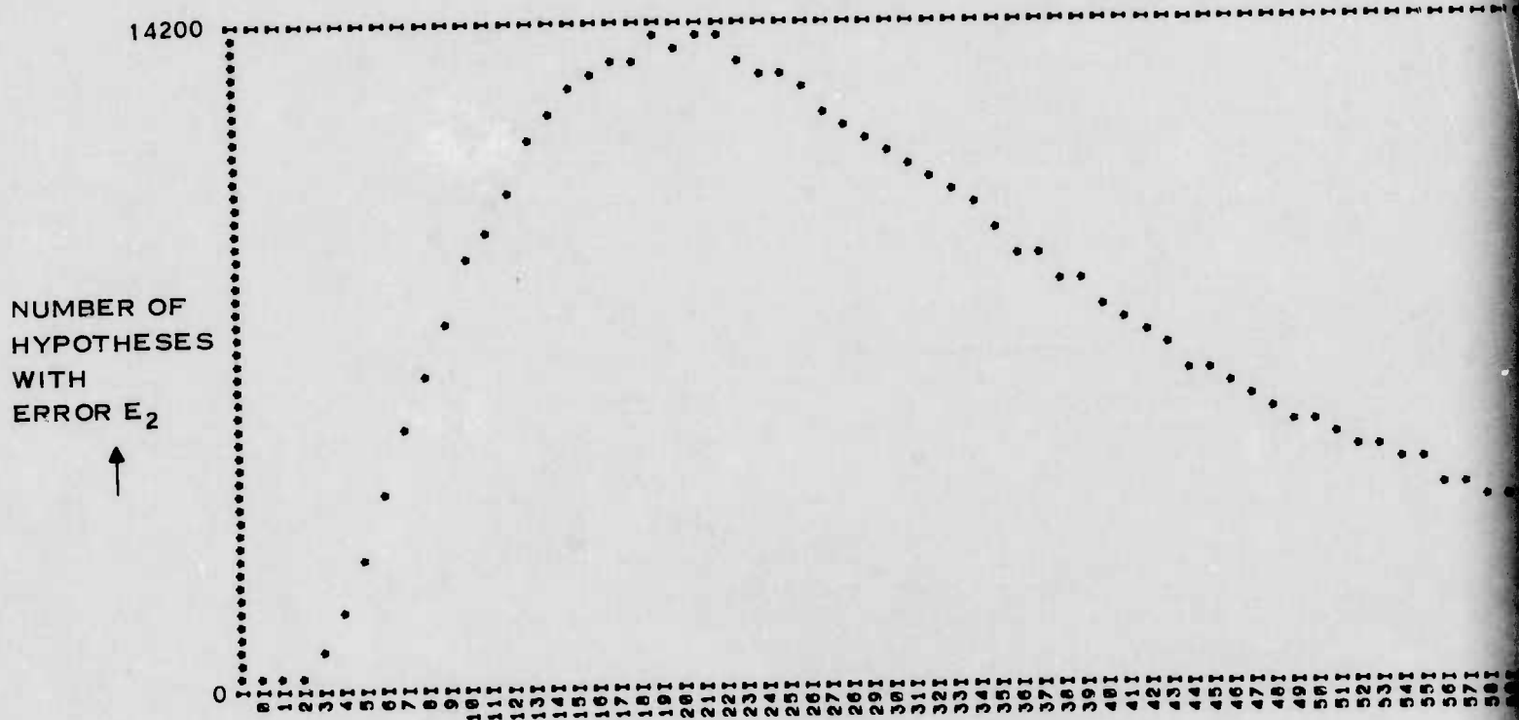
where

$$q(R, L_i) = \frac{M(R, L_i)}{\sum_{R \epsilon F} M(R, L_i)}$$

The final file F* is then

$$F^* = \{R \epsilon F : H(R) \leq H_o\}$$

where H_o is a fixed entropy threshold. Base 2 logarithms were used in computations; the maximum entropy possible is $\log_2 5 \doteq 2.322$.



191423

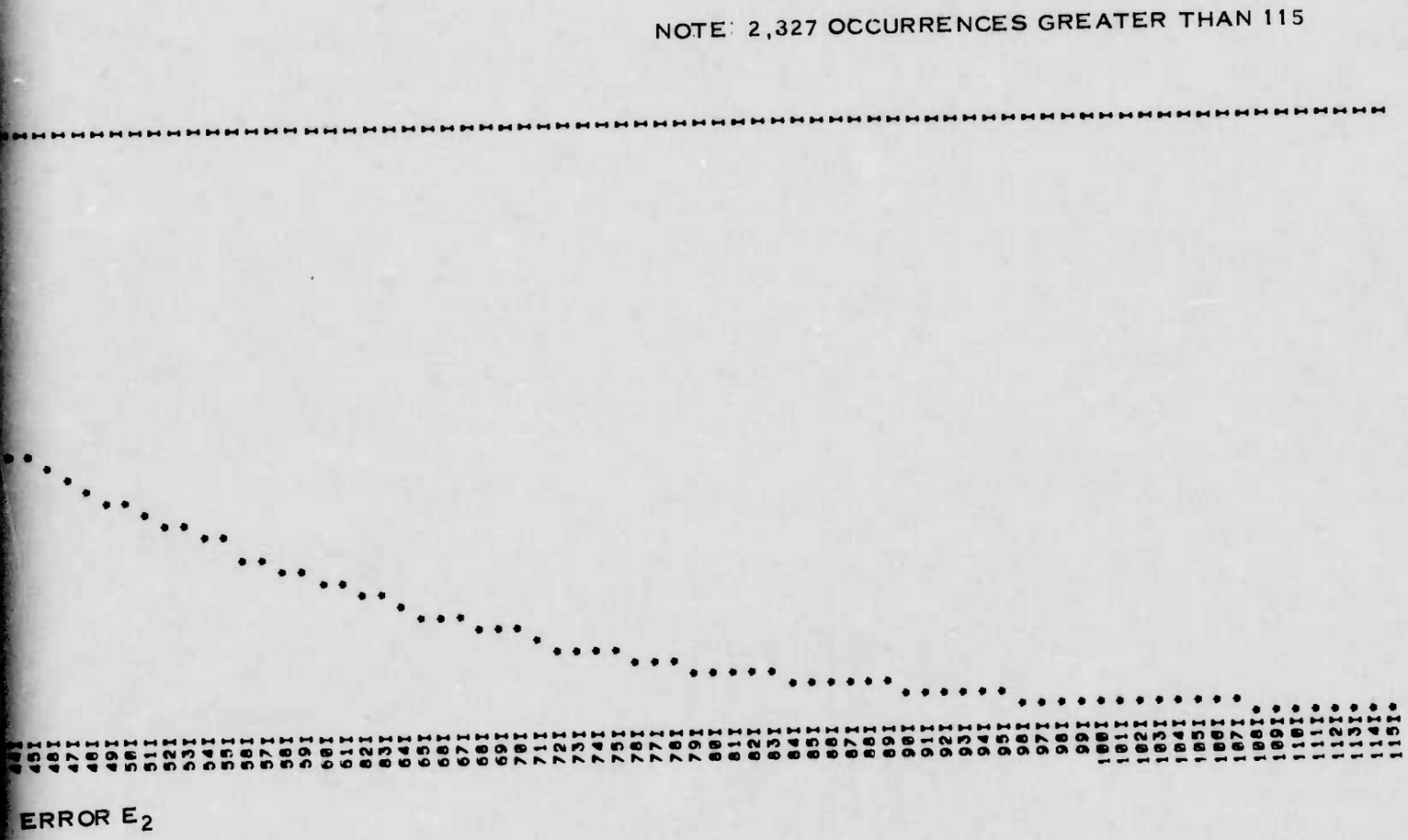
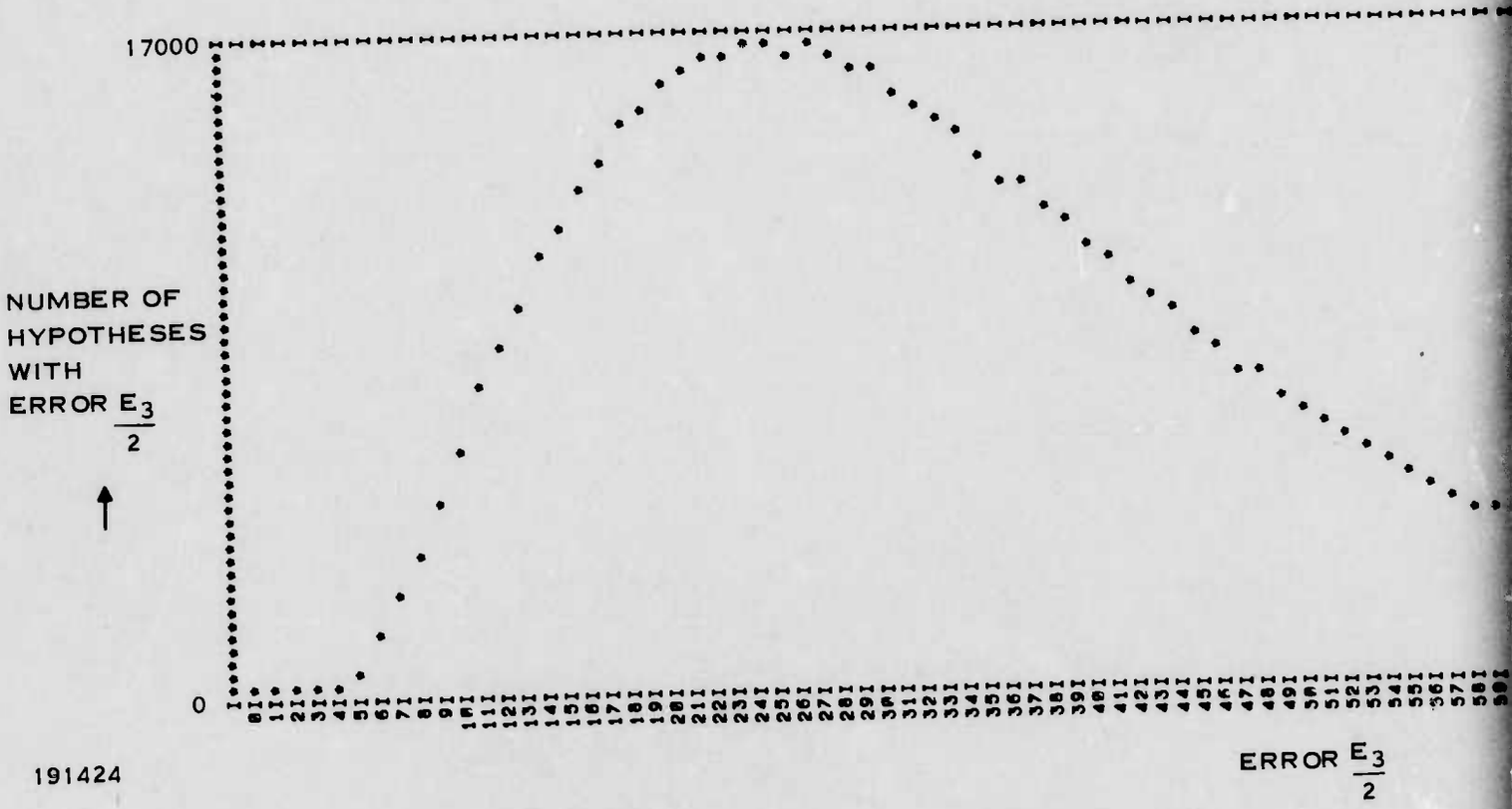
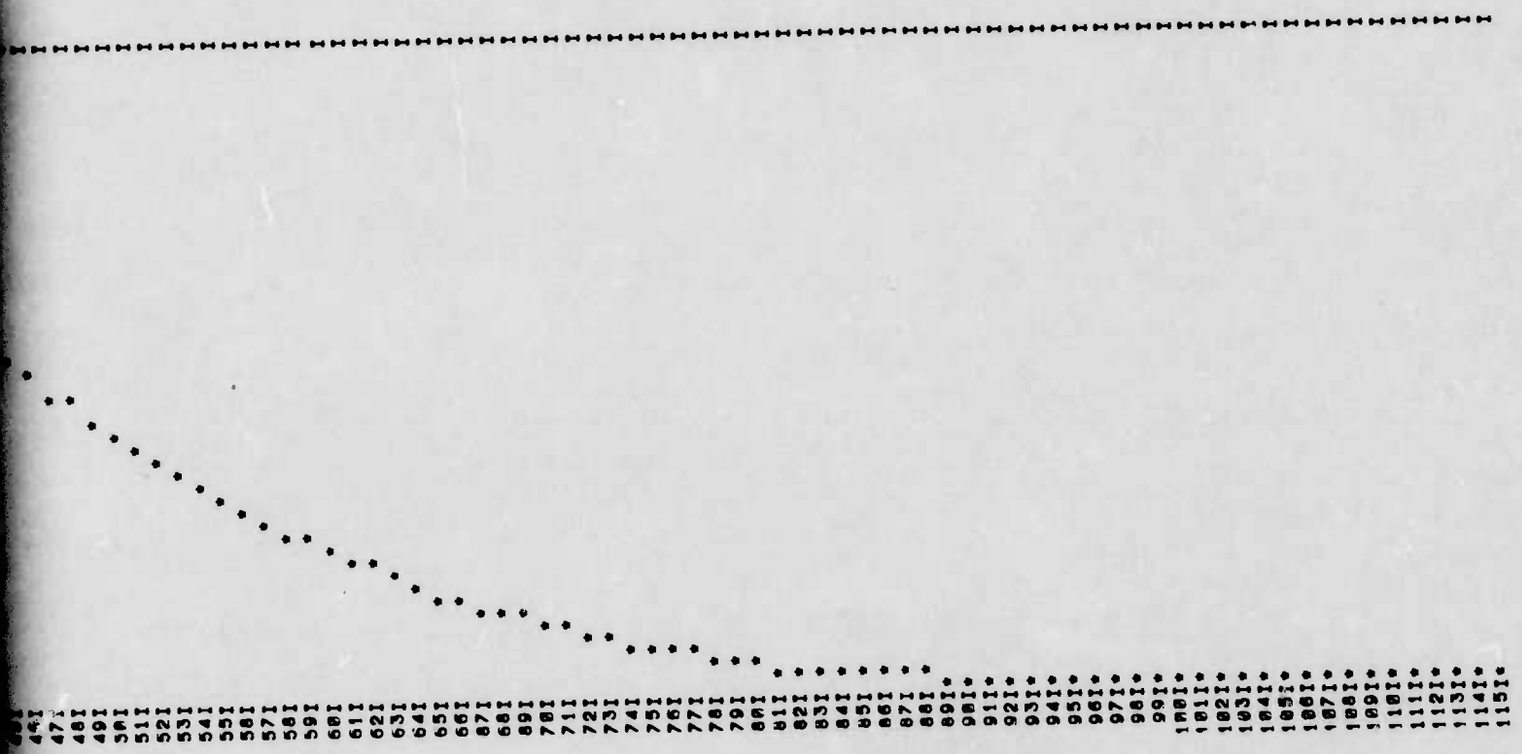


Figure 4. Frequency Distribution of Values of E_2

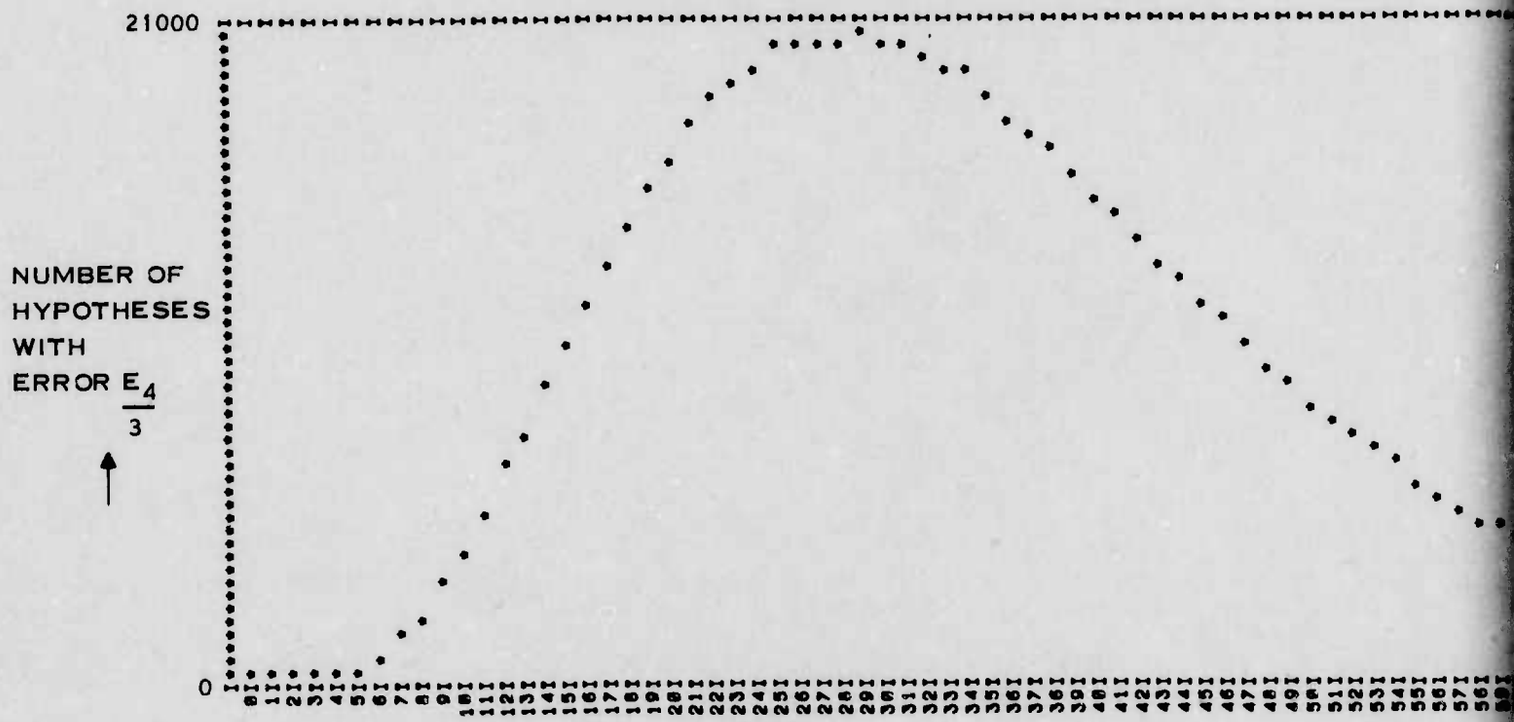

17000
0
NUMBER OF HYPOTHESES WITH ERROR $\frac{E_3}{2}$
ERROR $\frac{E_3}{2}$
191424

NOTE: 168 OCCURRENCES GREATER THAN 115



ERROR $\frac{E_3}{2}$

Figure 5. Frequency Distribution of Values of E_3 (Scaled)


21000
0
NUMBER OF HYPOTHESES WITH ERROR $\frac{E_4}{3}$
ERROR $\frac{E_4}{3}$

191425

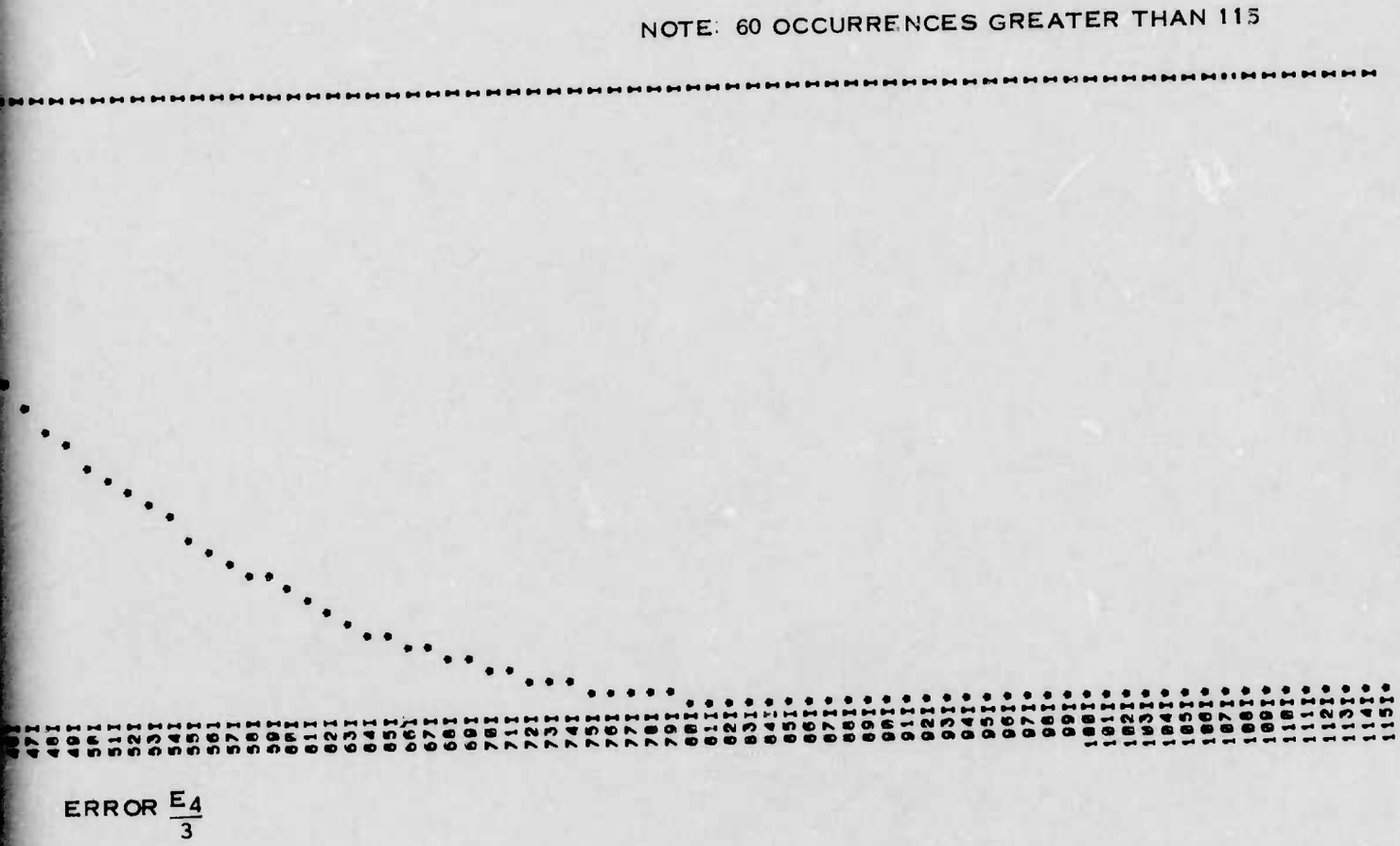
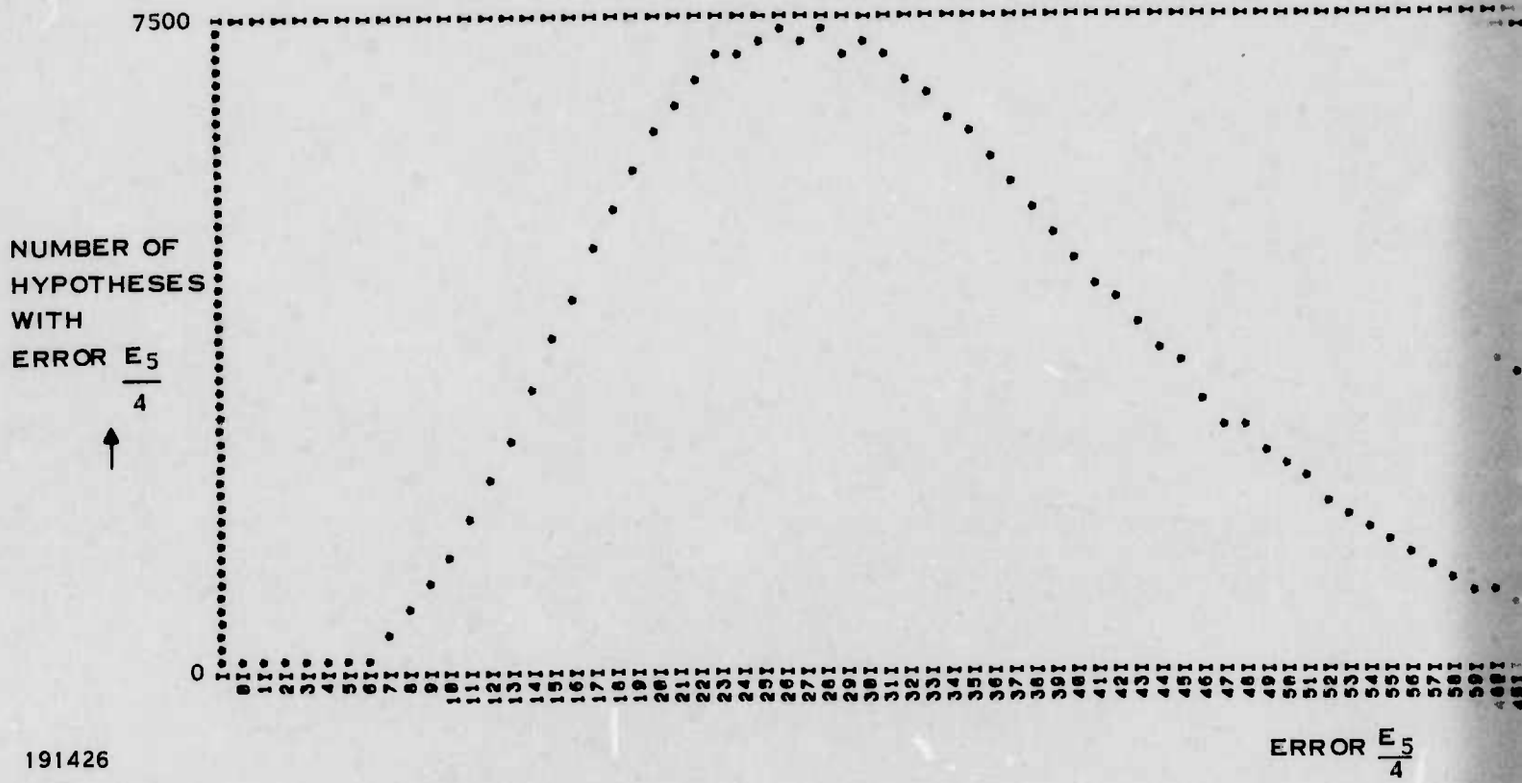


Figure 6. Frequency Distribution of Values of E_4 (Scaled)


7500
0
NUMBER OF HYPOTHESES WITH ERROR $\frac{E_5}{4}$
ERROR $\frac{E_5}{4}$

191426

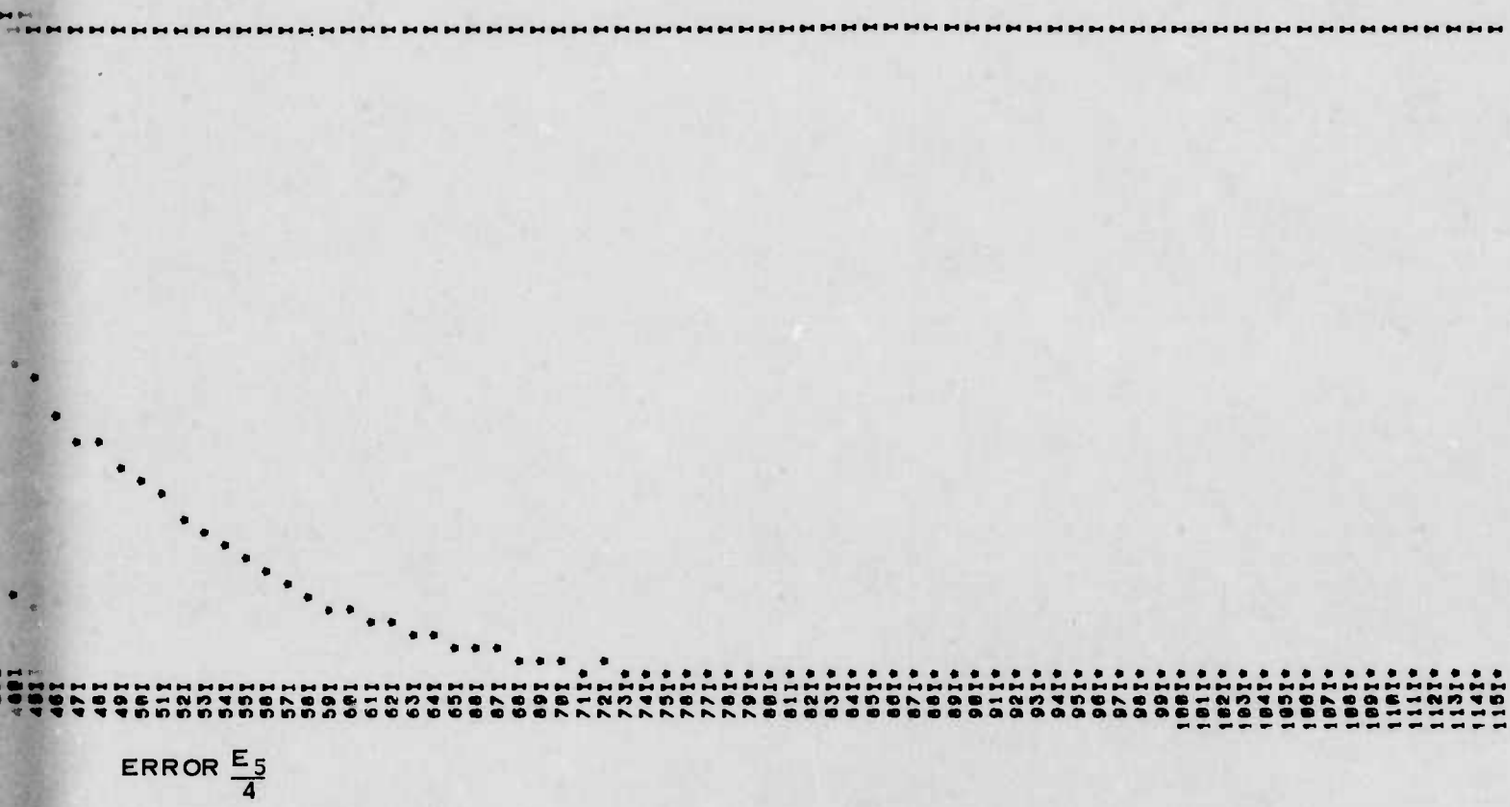


Figure 7. Frequency Distribution of Values of E_5 (Scaled)

SECTION IV
DECISION RULES

This section describes the decision rules used to classify data from test speakers. The general approach is to process a specified amount (90 seconds) of speech data to be classified (for example, from speaker S), detecting the occurrences of the references in F*. Let $R = \{R_1, R_2, \ldots, R_m\}$ denote the sequence of detected reference sequences of a specified length. R is taken to be the representation of the speech data from speaker S. Then the language likelihoods described above and the occurrence statistics of the sequence $\underline{R}$ are used to compute decision parameters which are used to classify the data as being from one of the five languages considered. Implementation of the decision strategies was carried out separately for each value of sequence length k, k = 2,3,4, and 5.

Let $p_i(R)$ denote the probability density function for the sequence $\underline{R}$, given that language L_i was spoken, i = 1,2,3,4,5. Letting P(L) denote the *a priori* probability for language L, the decision rule which is optimum in that it incurs the lowest possible probability of misclassification can be stated as: observe the sequence R and choose the language L_j for which

$$P(L_j)\, p_j(\underline{R}) \geq P(L_i)\, p_i(\underline{R}) \text{ for } i = 1,2,3,4,5$$

In practice, neither the *a priori* language probabilities nor the conditional sequence densities are explicitly known. Hence, approximations to the optimum rule must be used, and less than optimum results must be tolerated.

The basic strategy is to assume independence of the detected sequences of length k and then choose the language which maximizes the resultant expression for the log likelihood of the hypothesized languages given the observation of the data representation $\underline{R}$. Let $DF_1(S,L)$ denote the unnormalized decision function value computed for test speaker S and hypothesized language $L \in \{L_1, L_2, L_3, L_4, L_5\}$. Then

$$DF_1(S, L) = -\sum_{R \in \underline{R}} \log p(L|R)$$

where the summation is taken over reference sequences in the data representation $\underline{R}$ for speaker S. The following normalization is made. Define

$$DF_2(S, L) = \frac{DF_1(S, L)}{\sum_{S} DF_1(S, L)}$$

where the summation is over all 50 test speakers. Then let DF(S, L) denote the normalized decision function used to classify the test speakers, and define

$$DF(S, L) = \frac{DF_2(S, L)}{\left\{\sum_{i=1}^{5} \{DF_2(S, L_i)\}^2\right\}^{1/2}}$$

Decision rule A is to choose, for speaker S, the language $L \epsilon \{L_1, L_2, L_3, L_4, L_5\}$ for which DF(S,L) is smallest.

A second strategy is based on maximizing the correlation between the decision function values defined above for the test speaker and normalized decision function values. Let L_i and L_j denote actual language and hypothesized language, respectively, i, j = 1,2,3,4,5. Let

$$D_1(L_i, L_j) = \sum_{S \epsilon S_i} DF(S, L_j)$$

and then define

$$D(L_i, L_j) = \frac{D_1(L_i, L_j)}{\left\{\sum_{j=1}^{5} \{D_1(L_i, L_j)\}^2\right\}^{1/2}}$$

Decision rule B states: Compute the correlation

$$\rho^*(S, L_i) = \sum_{j=1}^{5} D(L_i, L_j)\, DF(S, L_j)$$

and choose, for speaker S, that language $L \epsilon \{L_1, L_2, L_3, L_4, L_5\}$ such that $\rho(S,L) = 1 - \rho^*(S,L)$ is smallest.

Each of these decision rules was implemented for each of four sequence lengths k = 2,3,4, and 5. Additional decision rules were used which were based on results for the four sequence lengths combined. To exhibit the dependence of the decision parameters on sequence length, k, define

$$DF'(S,L,k) = DF(S,L)$$

for sequence length k = 2,3,4, and 5. Then define

$$DF_c(S, L) = \sum_{k=2}^{5} DF'(S, L, k)$$

Decision rules A* and B* result from using DF_c instead of DF_2 and then making the succeeding computations in the same manner as for rules A and B, respectively.

SECTION V

CLASSIFICATION RESULTS

This section describes the classification results obtained using the various decision rules described in Section IV. Many classification experiments were performed to determine the exact structure of a classifier which performs well with the training data. The idea is to design the classifier using the training data and then classify the test speakers to estimate the probability of correct classification associated with that classifier. Variables that required specification were: (1) acceptance level for hypothesized sequences; (2) entropy threshold for selection of reference sequences with sufficient language specificity; (3) decision strategy; and (4) sequence length.

1. TRAINING DATA

Figures 8 through 12 show the numbers of errors from classification of the 50 training speakers as functions of the variables mentioned above. Figures 8 and 9 show (for decision rules A, A* and B, B*, respectively) the errors as a function of the acceptance level for acceptance of the hypothesis that a reference sequence has recurred. The cases for the various levels were described in Subsection III.3.d and were labeled 0, 1, 2, 3, 4, 5, as in the figures. These cases correspond to 100-, 50-, 25-, 12.5-, 6.25-, and 3.125-percent acceptance of all hypothesized sequences. These figures include results for each sequence length k = 2,3,4,5, and for each result plotted, the entropy threshold used was $H_o = 2.3$.

Figure 10 shows the errors incurred as a function of the entropy threshold for selection of reference sequences. This parameter determines the total numbers of references of each length remaining in the file used to compute decision function values. In each case shown, data from the 12.5-percent acceptance case was used.

Figure 11 shows the errors as a function of sequence length for the case: (1) $H_o = 2.3$, (2) 12.5-percent acceptance, and (3) decision rules A and B. This case was chosen because results for it were at least as good as for the other situations. For rule A of this same case, Figure 12 shows error as a function of the number of reference sequences remaining in the final reference file.

It is seen that, for the training speakers, Case 3 data (12.5-percent acceptance), an entropy threshold of $H_o = 2.3$, and decision rule A using sequences of length 4 yielded the best classification performance: 88-percent correct five-language classification.

2. TESTING DATA

Figures 13 and 14 show the numbers of errors resulting from classifying the 50 test speakers for various decision rules and values of the parameters. The parameters determined from the training data experiments provided the basis for choosing parameters for classification experiments with the testing data. Figure 13 shows performance as a function of acceptance level and Figure 14 shows performance as a function of entropy threshold. It is seen that Case 3 data and an entropy threshold of $H_o = 2.3$ yield the fewest errors provided that decision rule B was used with sequences of length 5. This choice yielded 62-percent correct five-language classification.

3. COMBINATION DECISION STRATEGY

A decision rule was formulated which used sequences of length k = 5 as well as single segments, as developed in the first phase of this study. During the first phase, values of a decision function $D^1(S,L)$ were determined (Table XI, reference 1) which were computed from the same test data. A decision function which combined the results of the two phases of this study was computed (for i = 1,2,3,4,5) to be:

$$D^c(S, L_i) = \frac{\rho(S, L_i)}{\sum_{S_j \epsilon \underline{S}_i} \rho(S_j, L_i)} + \frac{D^1(S, L_i)}{\sum_{S_j \epsilon \underline{S}_i} D^1(S_j, L_i)}$$

where $\rho(S,L_i)$ was computed for sequence length k = 5. Choosing the language which yielded the smallest value for $D^c(S,L)$ to classify the test speaker S yielded the confusion matrix shown in Figure 15. The overall classification accuracy resulting from this combination rule is 70 percent.

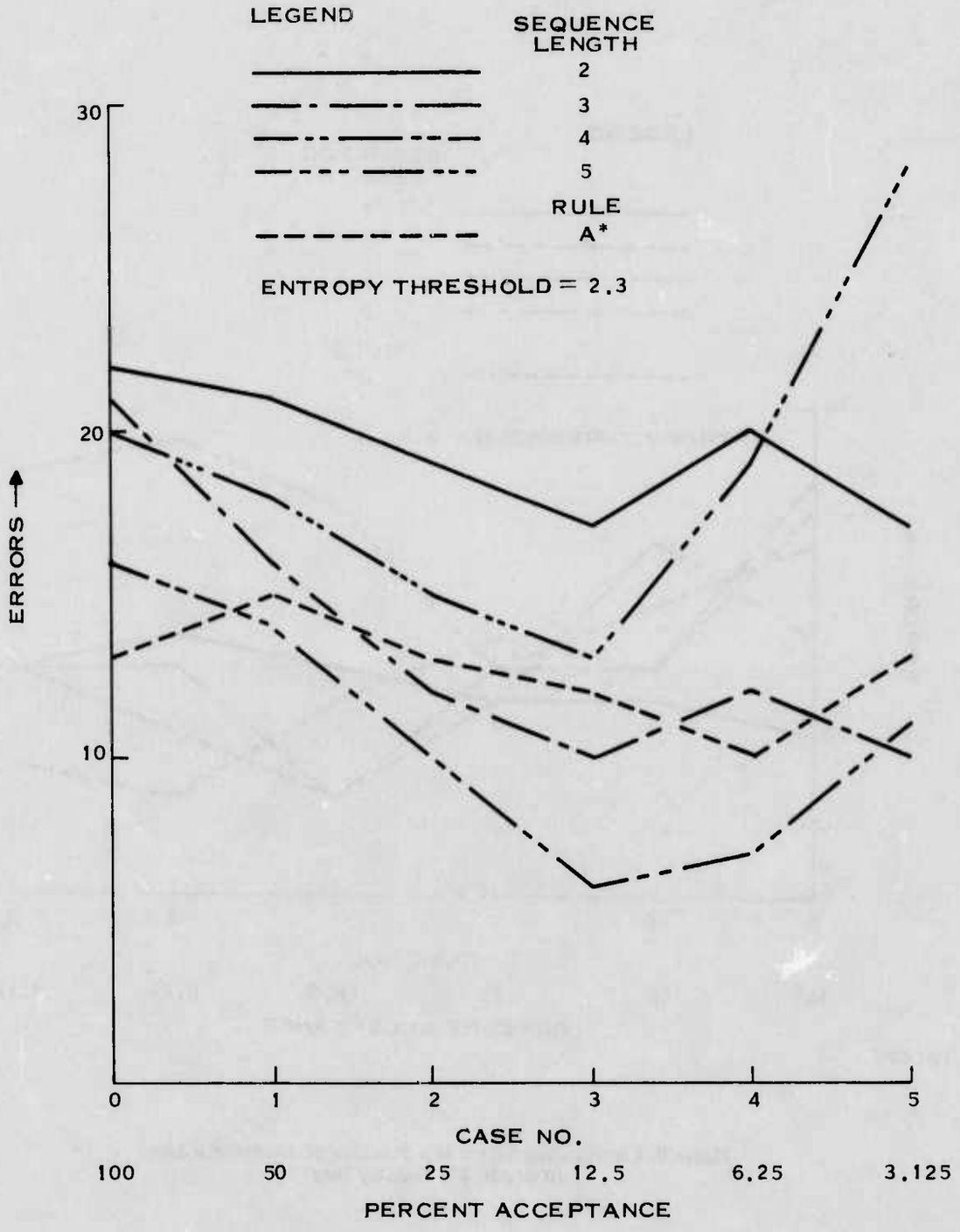


Figure 8. Classification Errors as a Function of Acceptance Level (Rules A, A*; Training Data)

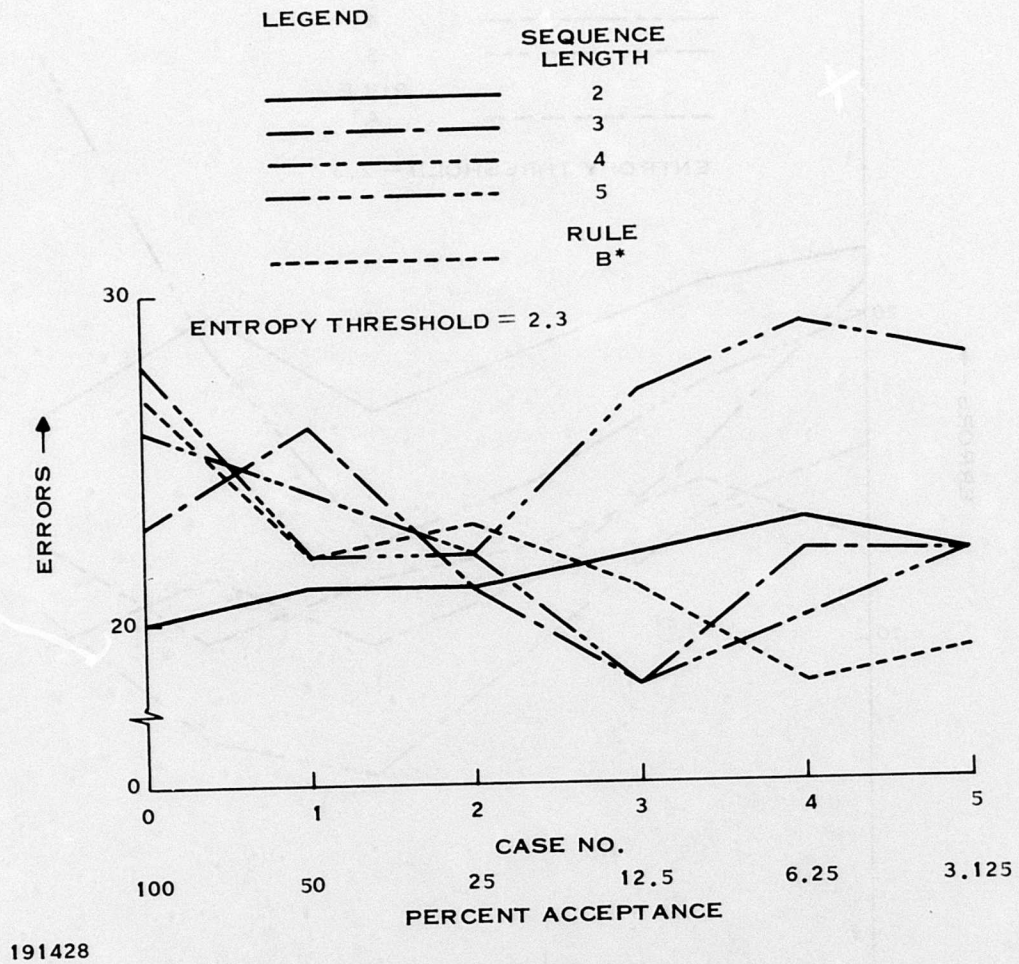


Figure 9. Classification Errors as a Function of Acceptance Level (Rules B, B*; Training Data)

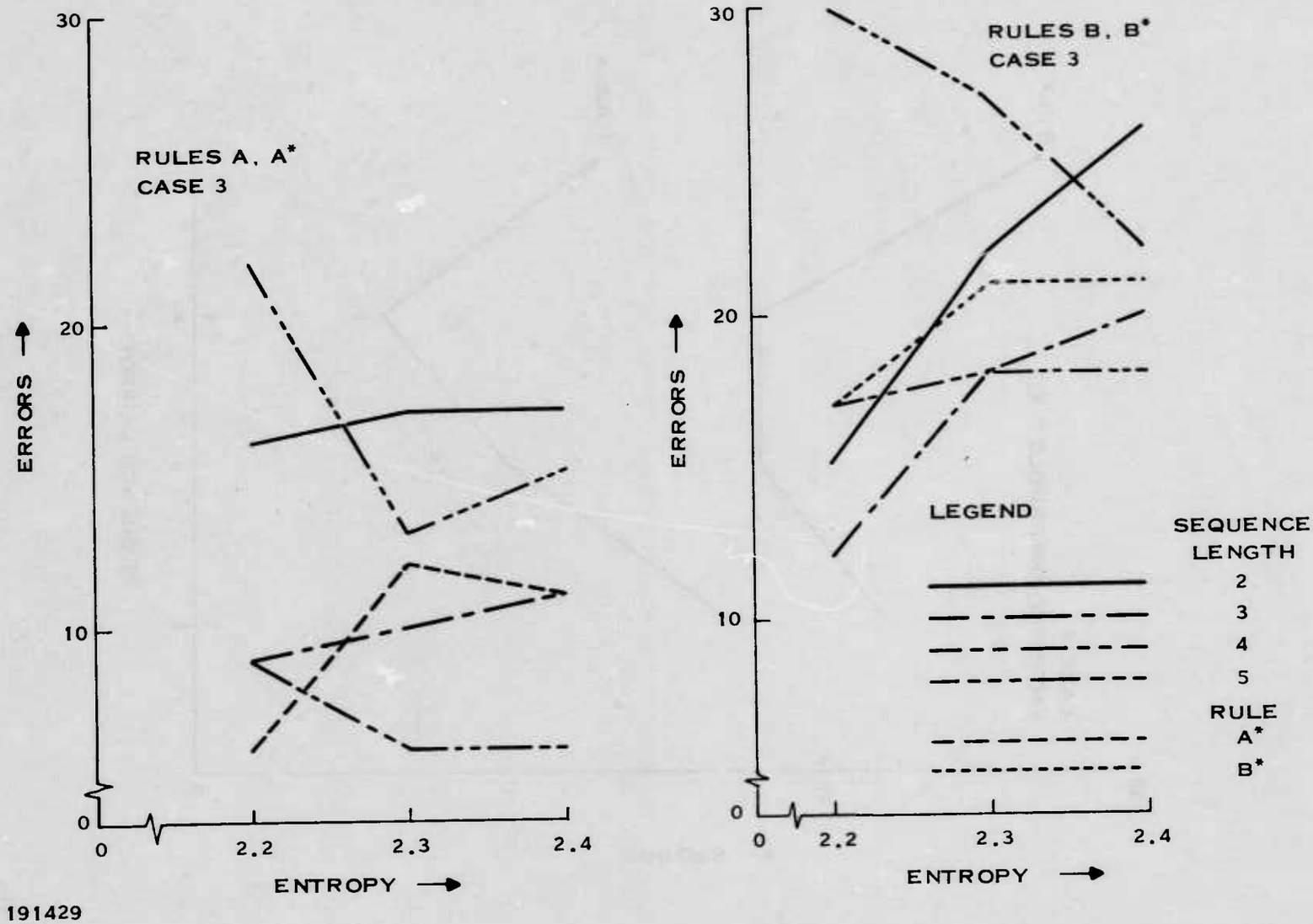
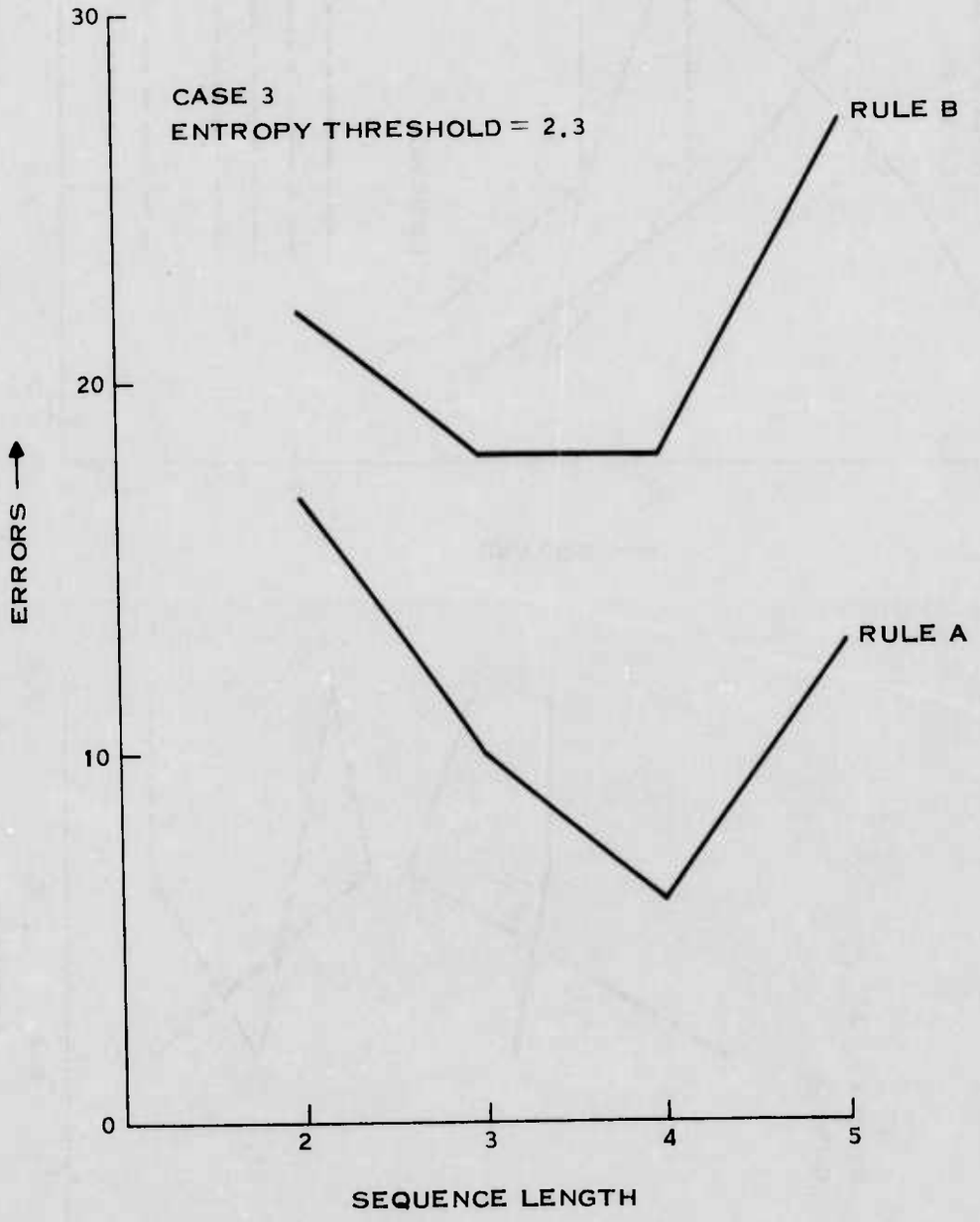
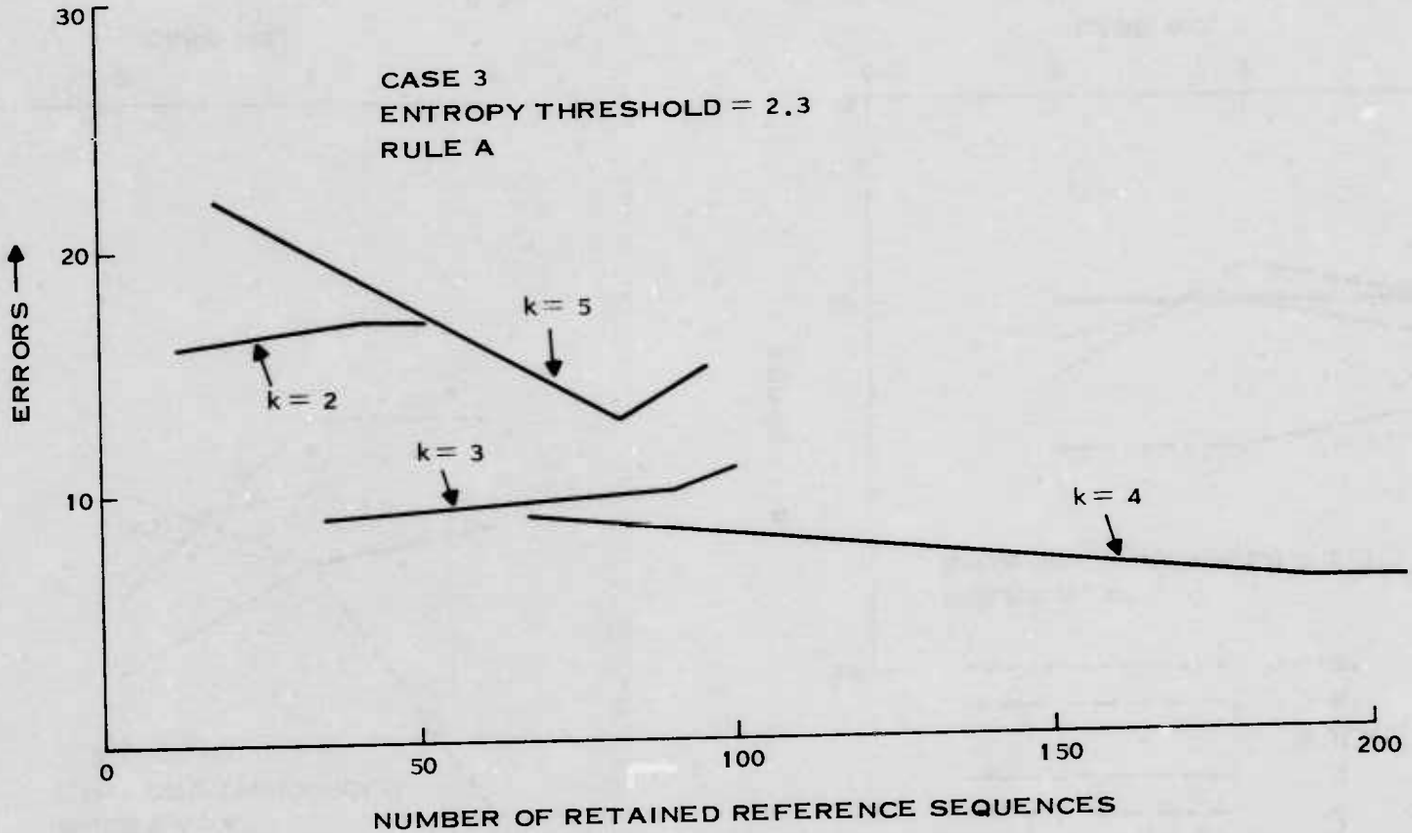


Figure 10. Classification Errors as a Function of Entropy Threshold (Training Data)



191430

Figure 11. Classification Errors as a Function of Sequence Length (Training Data)



191431

Figure 12. Classification Errors as a Function of Number of Retained References (Training Data)

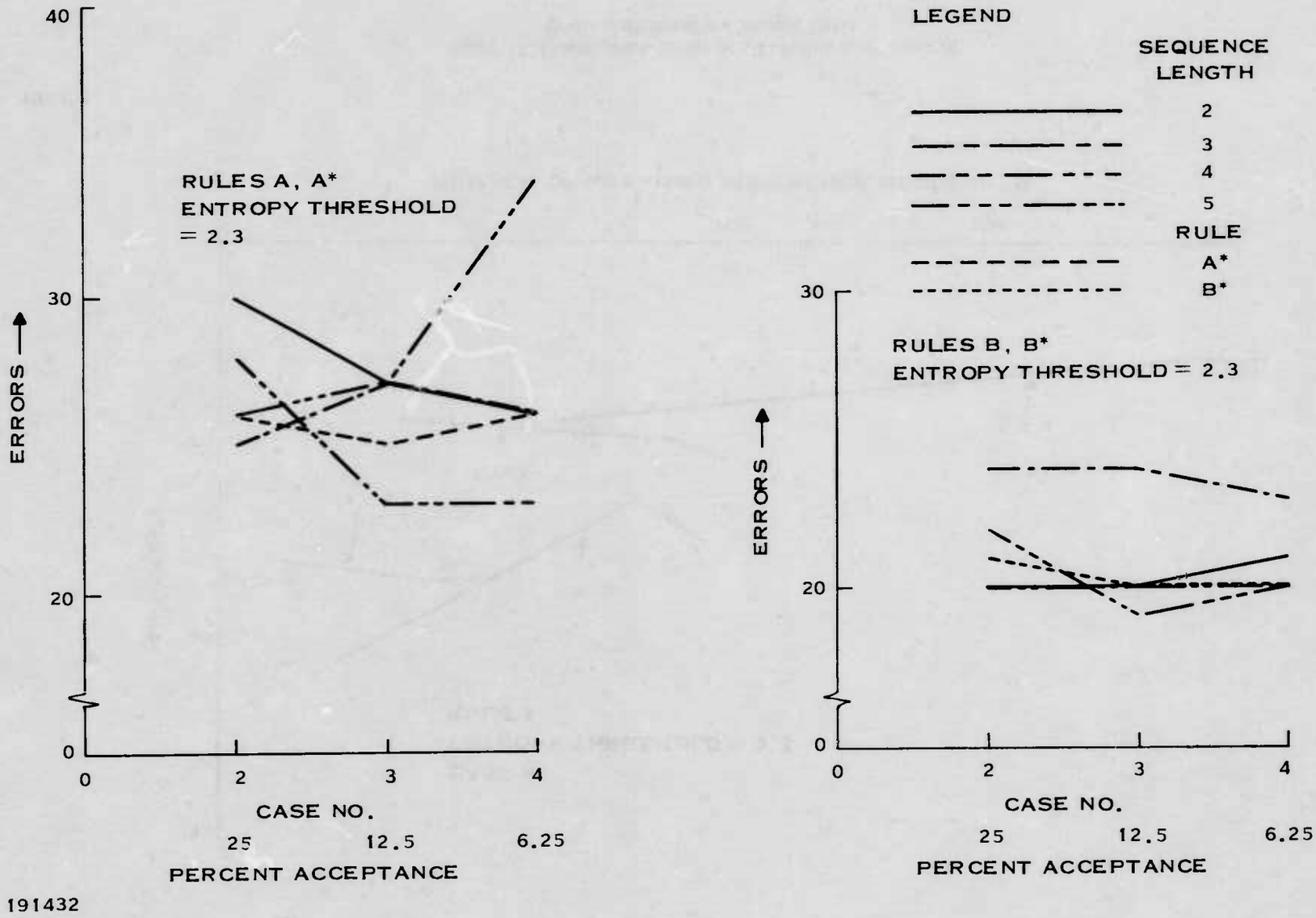


Figure 13. Classification Errors as a Function of Acceptance Level (Testing Data)

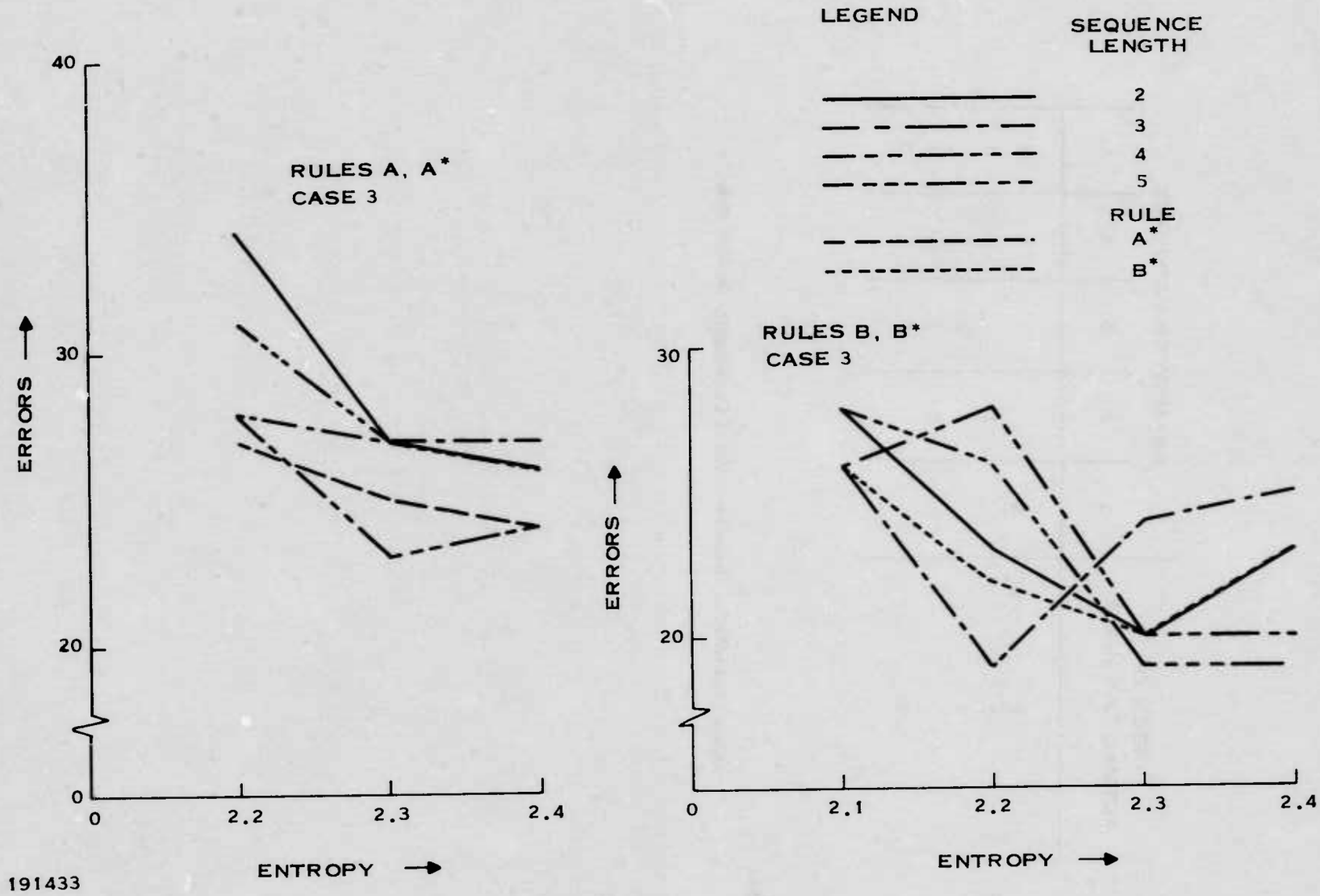


Figure 14. Classification Errors as a Function of Entropy Threshold (Testing Data)

INDEX OF ACTUAL LANGUAGE	ESTIMATED LANGUAGE				
	1	2	3	4	5
1	8			1	1
2		1		4	1
3			9	1	
4		2	1	10	1
5			2	1	7

191434

Figure 15. Confusion Matrix From Use of Combination Decision Rule

SECTION VI

CONCLUSIONS AND RECOMMENDATIONS

In this study language classification was based on sequences of 1, 2, 3, 4, and 5 phoneme-like sound segments. The approach taken treated each language identically, without special linguistic considerations. Time-frequency scanning was used to hypothesize time registration points and generate candidate reference sequences. Relationships among occurrence times, speech energy, and scanning errors were used to hypothesize the recurrence of reference sequences in input speech data. Classification was based on summed logarithms of the language likelihood estimates, given the occurrences of the reference sequences.

Sequences of length 4 performed best in classifying training speakers. For this best case, an entropy threshold of 2.3 provided the best rejection of sequences not having sufficient language specificity, and the acceptance threshold was set such that 12.5-percent of all hypothesized sequences were considered. In classifying the 50 training speakers, 88-percent correct five-language classification resulted.

A decision rule using sequences of length 5 in combination with sequences of length 1 (single segments) yielded best performance in classifying the test speakers, yielding 70-percent correct five-language classification. Again, the 2.3 entropy rejection level and 12.5-percent acceptance level for hypothesized sequences (as predicted from training data results) proved most useful when the independent test data was classified.

Speaker dependence proved to be a formidable obstacle in attaining good classification results. The same nine test speakers (18 percent of the test data base) were misclassified by both the decision rule B using sequences of length 5 and the decision rule using $D^1(S,L)$ for single segments. To reduce such speaker dependency, the following improvements are planned:

A voiced-data indicator and a pitch measure should be included in the spectral data representation.

Labeling of sequences as to basic sound type (e.g., stop, fricative, vowel, consonant) should provide better sequence classification as well as less speaker dependency.

To allow averaging the effects of individual speakers, separate representation of overall sequence data should be defined and used.

There should also be significant improvement in data processing throughput to allow more detailed understanding, analysis, and refinement of reference sequence files.

REFERENCE

1. R. Gary Leonard, and George R. Doddington, "Automatic Language Identification," Final Report, RADC-TR-74-200, August 1974 , (AD785397).

MISSION
of
Rome Air Development Center

RADC is the principal AFSC organization charged with planning and executing the USAF exploratory and advanced development programs for information sciences, intelligence, command, control and communications technology, products and services oriented to the needs of the USAF. Primary RADC mission areas are communications, electromagnetic guidance and control, surveillance of ground and aerospace objects, intelligence data collection and handling, information system technology, and electronic reliability, maintainability and compatibility. RADC has mission responsibility as assigned by AFSC for demonstration and acquisition of selected subsystems and systems in the intelligence, mapping, charting, command, control and communications areas.

Printed by
United States Air Force
Hanscom AFB, Mass. 01731