LEARNING MODELS AND REAL-TIME SPEECH RECOGNITION

by

Douglas G. Danforth, David R. Rogosa,

and Patrick Suppes

with an Appendix by Camille Bellissant

TECHNICAL REPORT NO. 223

January 15, 1974

PSYCHOLOGY AND EDUCATION SERIES

INSTITUTE FOR MATHEMATICAL STUDIES IN THE SOCIAL SCIENCES

STANFORD UNIVERSITY

STANFORD, CALIFORNIA

Learning Models and Real-time Speech Recognition*

Douglas G. Danforth, David R. Rogosa,
and Patrick Suppes

1. INTRODUCTION AND THEORY

In October of 1972, the decision was made at the Institute for Mathematical Studies in the Social Sciences (IMSSS) to use psychological learning models on the problem of computer recognition of human speech. In this investigation of speech recognition we used the standard home telephone as an inexpensive terminal for verbal communication in dealing with an educational curriculum, such as mathematics. In subsequent pages we describe mathematical learning models and some of their properties, their implementation as part of a speech recognition system, and a series of system experiments with children as subjects.

Speech recognition can be viewed as three separate processes: (a) the internal representation of each utterance, (b) the actual recognition process, and (c) the change of the internal representation upon the discovery of errors by the recognition process (learning). In our approach, the representation of each utterance is given by a vector of numbers U. These numbers are the digitized amplitudes and frequencies from three band-pass filters that take as input the analog signal from, say, a telephone (see Section 2). In this study we deal only with recognition of individual phrases, and consequently, each utterance may be normalized in time to a fixed length, 0.50 secs. Our recognition process utilizes what can be called the nearest neighbor approach. A metric (see below) is introduced into the space and a distance is calculated from the unknown utterance U to each of the members of a set of vectors

*This research was supported by National Science Foundation Grant EC 443X4.

{V} representing known phrases; the name of the V closest to U is assigned to U.

1.1 Theta Process

Upon the discovery that U was misclassified, the correct vector V is updated using the following learning model. Let $0 < \theta < 1$ be an arbitrary scalar parameter and let V be the old vector representing the word from which U is a sample. Then a new V representation can be constructed from a weighted average of U and the old V, namely,

$$V \leftarrow (1-\theta)*V + \theta*U. \qquad (1)$$

Note that as θ ranges from zero to one the new representation ranges from V to U. This model, called the theta process and patterned after psychological models developed by Bush and Mosteller (1955) and Estes and Suppes (1959), is one aspect of our learning approach to speech recognition. Let us now investigate some of the properties of this linear learning model. In what sense does V 'represent' a word? If U is considered a random sample from a population with mean vector M = EU, where E stands for expectation, then

$$EV \leftarrow (1-\theta)*EV + \theta*EU. \qquad (2)$$

If we initialize the representation V to the first-heard utterance from the population, then by a simple inductive argument we find that EV = M too, so that V is an unbiased estimate of the mean of the population to which U belongs. It is well known that the sample mean is also an unbiased estimate of the population mean. However, V has the property of giving greater weight to recent utterances than to earlier ones. This responsiveness of V is useful in providing a more accurate representation of the speaker's current pattern of speaking.

It is of interest to consider the distance of a sample U of the population to its representation vector V so as to determine the likelihood of correct classification. Let d(U,V) be this distance and Ed(U,V) its expected value. If we assume a Euclidean metric and independent, identically distributed (i.i.d.) random samples U, it can easily be shown that this distance on the nth trial is given by

$$Ed(U,V) \quad = \quad 2\,\frac{1 + (1-\theta)^{2n-1}}{1 + (1-\theta)} * Ed(U,M), \qquad n=1,2,\ldots \tag{3}$$

where Ed(U,M) is unknown but independent of θ and trial number. Thus we have an expression for the expected distance between a member of the population and its representation vector V. Figure 1 shows explicitly the functional form of Ed(U,V)/Ed(U,M) for the cases where n=10 and n=50.

Insert Figure 1 about here

Equation 3 gives the expected distance as a function of n and θ. The minimum of Equation 3 may be obtained by setting its derivative equal to zero. By fixing n and solving for θ in this expression, the values presented in Table 1 were obtained. Note that n set to zero signifies V = U initially. This table is considered later in the experimental sections with regard to the error rate of classification.

Insert Table 1 about here

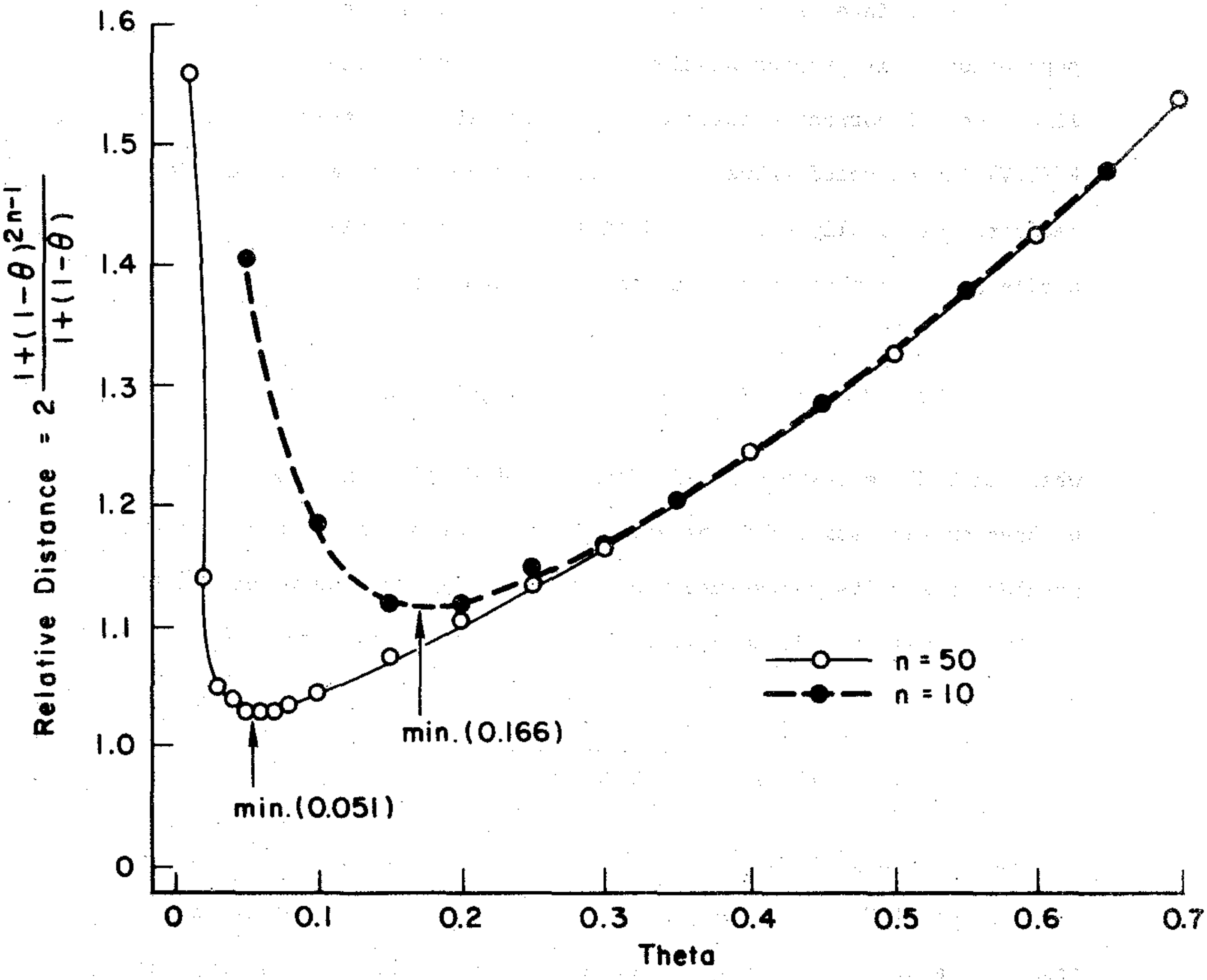


Fig. 1. The above plot displays the existence of very pronounced minima at θ=0.166 for n=10 and at θ=0.051 for n=50. Minima such as these occur for each trial number n. When θ assumes one of these minimizing values, it is reasonable to believe that the probability of an utterance being correctly classified as V is maximized.

TABLE 1

Values of Theta Which Minimize the Expected Distance

as a Function of n

n	θ	
1	1.000	
2	0.472	
3	0.387	
5	0.252	
10	0.166	(see experiment 2A)
15	0.126	
20	0.103	
25	0.087	
30	0.076	
35	0.068	
40	0.061	
45	0.056	
50	0.051	(see experiment 1)

1.2 Delta Process

The theta process is essentially an estimate of the first moment of the population. In the standard problem of statistical classification (Anderson, 1958), estimates of the covariance matrix are necessary to determine a hyperplane separating two populations. In order to avoid the inversion of a full covariance matrix, which is necessary with the classical Baysian procedure, one may use other less precise but, computationally more efficient techniques. One of these, which we call the delta process, estimates the variances of the utterance components. Let δ (delta) be a parameter that lies in the interval 0,1 , then S^2 given by the learning equation

$$S^2 \leftarrow (1-\delta)*S^2 + \delta*(U-V)^2 \qquad (4)$$

is an estimate of the component variances, where U,V are as before. Using the two quantities V and S^2, we may calculate a 'distance' between the utterance U and a representation vector V by

$$D(U,V) = (U-V)^T W\,(U-V), \qquad (T=\text{Transpose}) \qquad (5)$$

where

$$W = \frac{A}{TrA}, \qquad (Tr=\text{Trace}) \qquad (6)$$

and

$$A = \text{diag}(S^2)^{-1}, \ \text{diag}(S^2) = \begin{pmatrix} S_1^2 & & 0 \\ & S_2^2 & \\ & & \cdot \\ 0 & & S_n^2 \end{pmatrix}, \qquad (7)$$

which differs from the Euclidean distance d(U,V) by the replacement of I (the identity matrix) by W. Notice that components with high variability are weighted less than those with low variability.

1.3 Beta Process

Alternatively, we may introduce the concept of a strength associated with each component of V and then increase or decrease its value depending upon whether that component correctly or incorrectly classifies an utterance. Let L be a vector of strengths associated with V. Then L_i can be changed by multiplying by a quantity β_i (Beta) such that

$$L_i \leftarrow \beta_i * L_i. \tag{8}$$

Thus, $\beta_i > 1$ if i is a good component and $\beta_i = 1$ if it is bad, (Eq. 10). The weights subsequently associated with the components of V are related to the strengths through normalization, namely

$$W = \frac{A}{TrA} \tag{6'}$$

where

$$A = (\text{diag } L). \tag{9}$$

Again the distance between an utterance and a representation vector V is given by

$$D(U,V) = (U-V)^T W (U-V). \tag{5'}$$

A good component is defined when an error in classification has occurred Let V' be the incorrectly chosen representation vector and V the true vector with which U should be identified. Then component i is good if

$$(U-V)_i W_{ii} (U-V)_i < (U-V')_i W'_{ii} (U-V')_i \tag{10}$$

and bad otherwise. Changing the strengths by Equation 8 is Luce's beta process, which has been studied extensively in Lamperti and Suppes (1960). We call the combined processes (theta,delta) the delta model and those of (theta,beta) the beta model.

1.4 Internal vs External Learning Models

Our use of the delta and beta models is at variance with what is usually done in the psychological investigation of human learning. A task is presented to subjects, and a mean learning curve is obtained by measuring the average number of correct responses as a function of the presentations of the task (trial number). A theoretical model is then proposed as a possible explanation for this correct response curve, and the parameters of the model are estimated from the data. We may consider such models 'external' models. In contrast, we specify explicitly the internal response processes. Consequently, the delta and beta models, as used here, may be considered 'internal' models. The theoretical link between the internal-external responses of the machine is suggested in Sections 3.4 and 4.3 through the comparison of the minimum expected distance of an utterance to its representation vector and the measured error rate of experiments 1 and 2. Further theoretical investigation of this link is underway. Section 4.2 and 4.3 discuss the application of an external model to the learning curves of experiment 2A.

2. IMPLEMENTATION

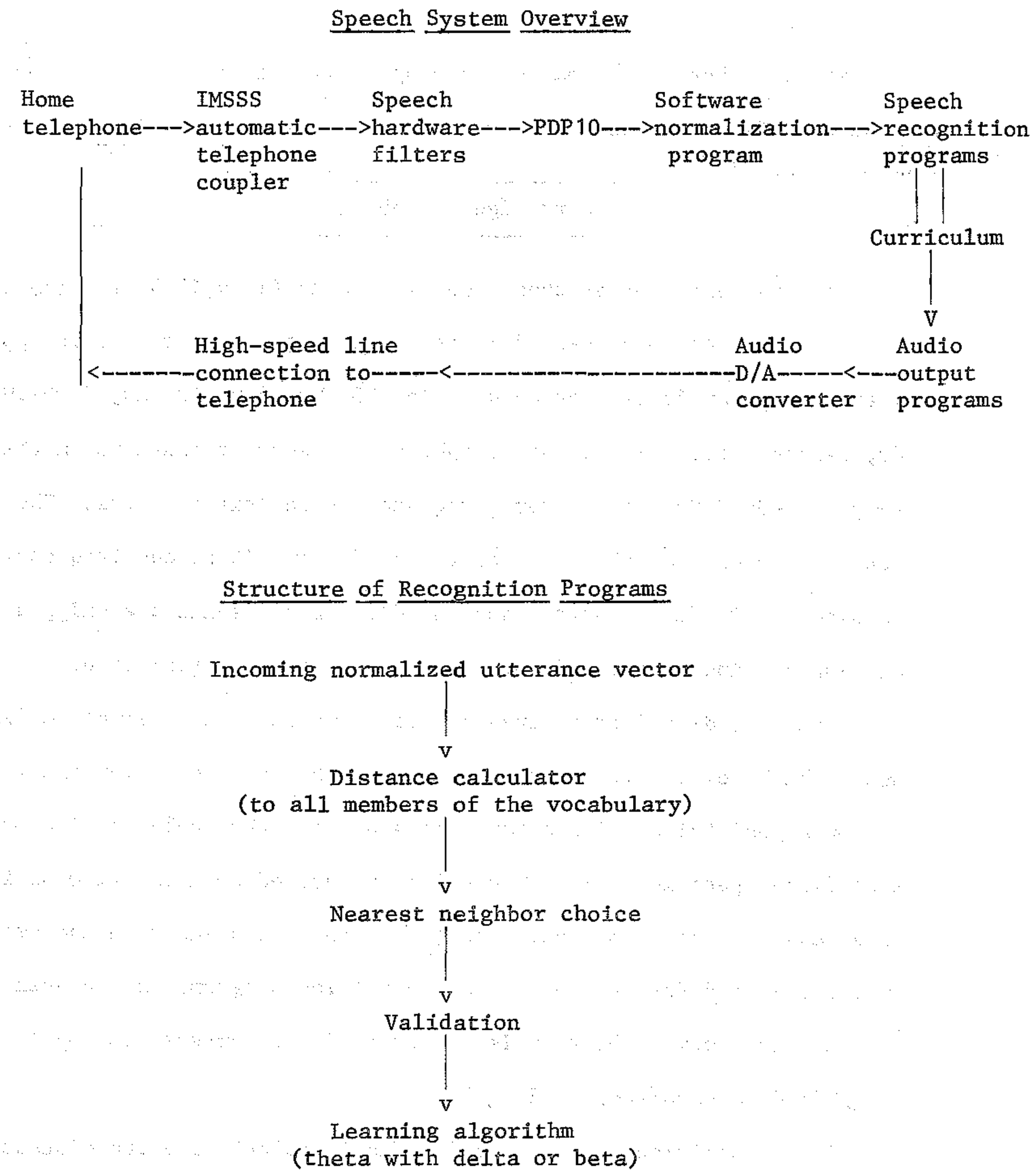
An overview of our speech recognition system is presented pictorially in Figure 2. A call placed from a standard home telephone to an

Institute number is automatically coupled with an Institute high-speed line that feeds the analog signal to our hardware filters. These filters are patterned after those used in Vicens (1969,1970) and consist of three solid-state band-pass filters whose ranges were chosen to approximate the human formant structure--150-900 Hz, 900-2100 Hz, 2100-5000 Hz, respectively. Since the telephone frequency response is in the range 300-3000 Hz, our filters adequately span this interval. The output from each of these detectors then is amplitude and frequency sampled at 10 msec intervals and the digitized results are shipped by high-speed line to our PDP-10. This is all done in real time.

Insert Figure 2 about here

The raw, digitized utterance data flows into an internal buffer until the hardware stops transmitting, which occurs whenever the input analog signal falls below a hardware specified threshold for longer than a hardware specified time. The buffer is dumped when the flow of input data ceases. The dumped data are then reformatted and time and amplitude normalized for return to the recognition programs in a convenient standardized form. The form is a vector of 300 numbers, (3 amp + 3 freq)*(100 samples/sec)*(1/2 sec).

The recognition process simply entails calculating the distance from utterance vector U to each representation vector V of the vocabulary, that is, calculating the weighted sum of squares of component differences. The word with the minimum distance is deemed the best choice. The recognition rate is such that some 30 words per CPU second



Speech System Overview

Home telephone
IMSSS automatic telephone coupler
Speech hardware filters
PDP10
Software normalization program
Speech recognition programs
Curriculum
Audio output programs
Audio D/A converter
High-speed line connection to telephone

Structure of Recognition Programs

Incoming normalized utterance vector
Distance calculator
(to all members of the vocabulary)
Nearest neighbor choice
Validation
Learning algorithm
(theta with delta or beta)

Fig. 2. Overview of speech recognition system.

can be compared. In experiment 2B, where a 14 word vocabulary is used, actual recognition times in a time-sharing environment and optimal recognition times are comparable (about 1/2 sec).

Changes of the internal representation of the word spoken are accomplished by the learning algorithms based on the theory previously described. Specifically, this entails modifying each component of the representation vector V and its associated strength vector (S^2 or L).

The programming requirements of the two models are quite minimal. The programs are written in SAIL (Stanford Artificial Intelligence Language), which is a superset of ALGOL. The full curriculum of experiment 2B occupies, when running, only about 35K of core memory including the child's state vector (see Secs. 4 and 5), the recognition algorithms, and the audio output routines.

The production of spoken output is presently accomplished by retrieving digitized representations of the words stored on magnetic disk and by software regeneration of the analog signal. Again this audio process is executed in real time. Consequently, the interchange between student and computer is sufficiently fluent for smooth verbal communication with the educational curriculum.

3. EXPERIMENT 1

3.1 Description

As a first quick test of the models, two highly confused utterances, the letters B and D, were chosen. Fifty utterances of each letter were spoken into a high quality crystal microphone and recorded on disk in their digital form, after having passed through our hardware filters. These utterances were then cycled 10 times, in their original order, through the delta and beta models.

3.2 Delta model

The parameters θ and δ for the delta model ranged in the interval $0.1, 1.0$ and $0.1, 0.4$, respectively. Larger intervals were not used as the basic structure of the delta model was revealed in this range. Table 2 gives the results of the percentage of correct classifications (PCC) for the grid space. Note that under these somewhat artifical conditions the delta model performed well with a regular structure and a recognition rate of 96 percent at $\theta=0.1$ and $\delta=0.1$.

Insert Table 2 about here

3.3 Beta Model

Table 3 shows results of the beta model using the same data. Note, at least in the preliminary test, a somewhat poorer performance (81 percent at $\theta=0.1$ and $\beta=1.1$) with less regularity of structure than the delta model.

Insert Table 3 about here

3.4 Theta Process

A different, but similar, set of data (50 utterances each of B and D) was used to examine the theta process by itself. Again the data were

TABLE 2

Percentage of Correct Classification,

Delta Model

Theta										
.40	61	58	53	+	+	+	+	+	+	+
.30	70	65	65	59	56	58	+	55	53	57
.20	80	75	68	66	67	63	52	56	+	57
.10	96	95	83	78	75	71	61	45	37	57
	.10	.20	.30	.40	.50	.60	.70	.80	.90	1.00

Delta ->

+ not processed.

TABLE 3

Percentage of Correct Classifications,

Beta Model

	Beta										
Theta	1.0	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
1.00	66	71	70	70	69	+	65	62	58	64	61
.90	59	62	58	57	55	+	54	54	54	53	53
.80	58	63	61	59	60	+	63	57	53	57	67
.70	64	66	63	75	57	59	74	56	61	54	71
.60	69	69	71	56	58	55	56	68	53	58	60
.50	64	73	65	58	58	62	58	59	59	53	66
.40	73	72	72	62	64	60	66	54	58	70	75
.30	70	74	66	69	78	68	64	78	53	53	76
.20	73	76	62	64	70	62	63	79	66	77	77
.10	76	81	78	63	62	58	79	81	79	79	79

+ not processed.

cycled 10 times using the beta model with β set to one (i.e., no change of strengths), and allowing θ to vary from 0 to .1 in steps of .01 and from .1 to 1 in steps of .1 . The form of the curve in Figure 3 and the occurrence of the minimum at .045 for θ, after 50 distinct trials, correspond closely to the prediction of Figure 1 for the minimum distance to the representation vector; however, the similarity of error rate and expected distance is blurred by the fact that the 50 distinct utterances were presented ten times to the learning model. It can be after 50 distinct trials, correspond closely to the prediction of Figure 1 for the minimum distance to the representation vector; however, the similarity of error rate and expected distance is blurred by the fact that the 50 distinct utterances were presented ten times to the learning model. It can be considered, however, that each cycle is a sequence of 50 distinct utterances differing only in the starting configuration.

Insert Figure 3 about here

This preliminary experiment shows promise for the learning-model approach to speech recognition (delta model 96 percent) and indicates that the theta process is amenable to relatively simple analysis (error rate and expected distance similarity).

4. EXPERIMENT 2, PART A

4.1 Description

In an effort to provide a practical test of the two models under actual operating conditions of telephone transmission and reception, we

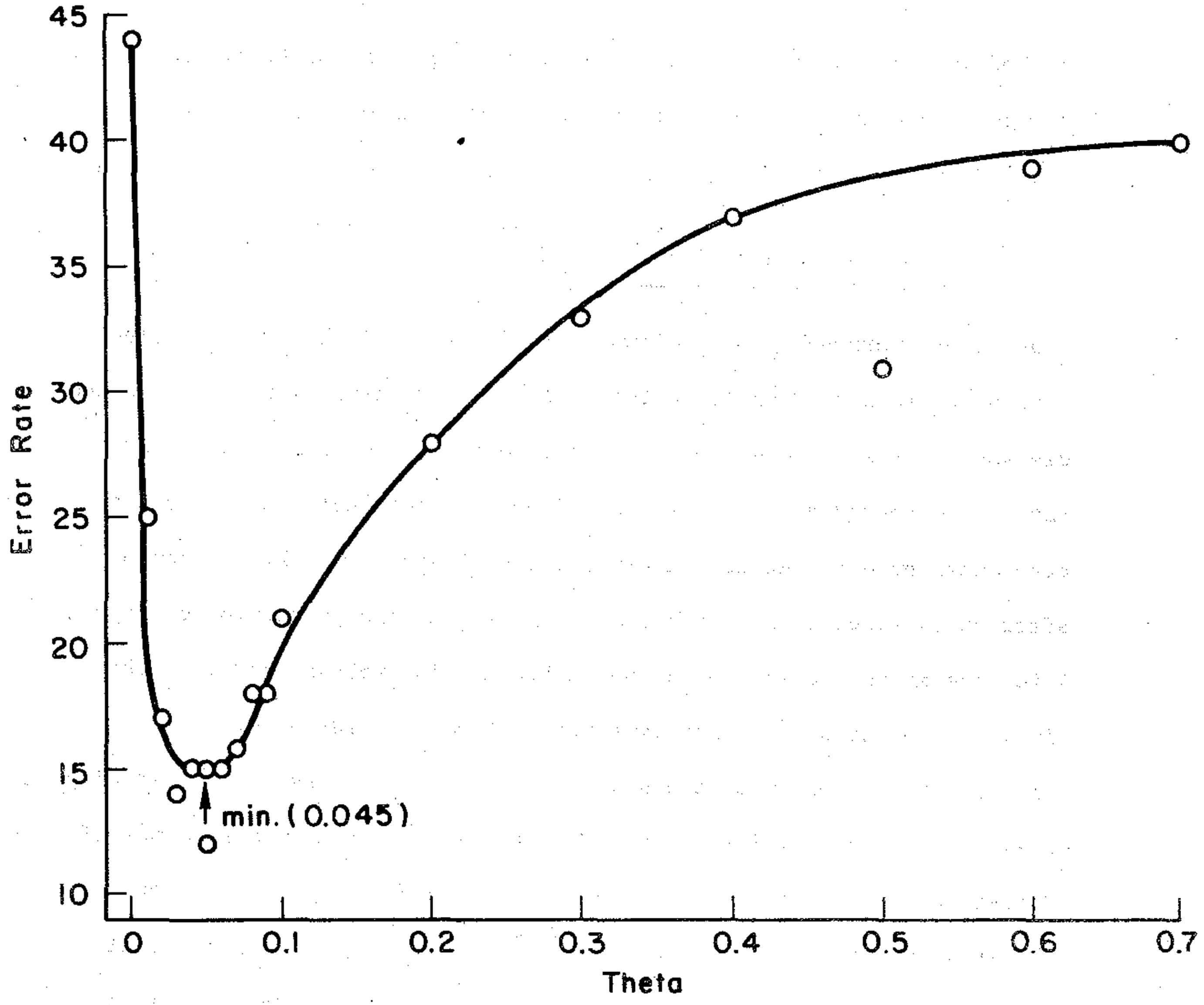


Fig. 3. Error rate curve for the theta process.

designed and executed an experiment of two parts (A and B). In A we acquired a data base of 14 children's voices spoken over the local Palo Alto telephone system. The telephone arrangement, described in Section 2, entailed calling a local Palo Alto number connected to the Institute from a university extension. The children, 3 girls and 11 boys, ranged in age from 6 to 13 years. A 14-word vocabulary (consisting of the digits 0-9 and command words yes, no, repeat, and stop) was chosen for compatibility with an elementary mathematics curriculum, Dial-A-Drill (Computer Curriculum Corp., 1971). In the experiment the vocabulary was presented sequentially on a cathode-ray tube terminal and was repeated by the child into the telephone for a total of 11 repetitions of each word. The time and amplitude normalized form of each utterance was recorded on magnetic disk.

For the analysis, the data were sequentially presented to the beta and delta learning models in a machine representation of actual speaking conditions. A parameter grid space was spanned for each model (Figs. 6, 7) and the recognition rates were examined to determine the optimal parameter settings.

4.2 Learning Curves of Correct Classification

We can represent the results of this experiment by learning curves for both the delta and beta models. To form each learning curve, we combined each of the 14 subjects and their 14 responses per trial to form a learning curve with ten points, with each point representing 196 subject-items on that particular trial. Since we have 11 repetitions of the vocabulary for each child, each learning curve has ten data points.

As an illustration that actual machine learning is taking place, we examine these curves in the context of mathematical learning theory. In order to avoid imposing a specific 'external model' on the learning process, we examine the mean learning curve, since the same mean learning curve can be generated from a wide variety of models. When we define the asymptotic response probability P(correct)= π (as n goes to infinity), a 'guessing' parameter p_0, and a learning parameter $0<X<1$ we obtain the mean learning curve

$$P(\text{correct on trial } n) = \pi - (\pi - p_0)X^{n-1} . \qquad (11)$$

In Figure 4, we see that the learning curve for the delta model attains a value of about 95 percent correct responses. The deviations

Insert Figure 4 about here

of recognition rates from the average across children are indicated by the ± one standard deviation error bars for each trial number. The shape of the learning curve indicates that at least five repetitions of the vocabulary are necessary for a high recognition rate. As shown in Figure 5, the learning curve for the beta model reaches 91 percent correct responses, which is lower than the delta model.

Insert Figure 5 about here

4.3 Regression on the Learning Curves

Since we are considering the general mean learning curve independent of a specific learning model, we will not estimate the parameters

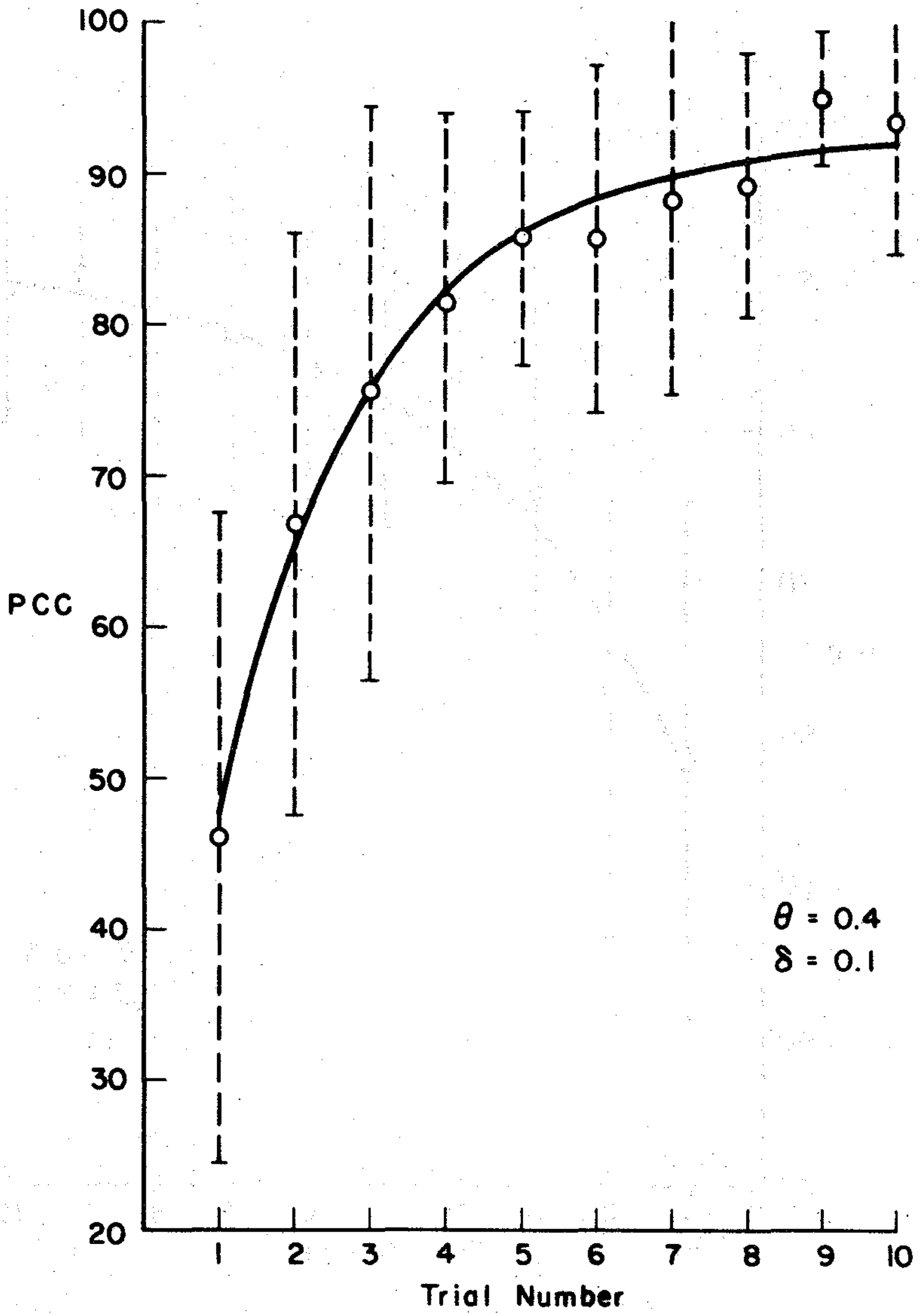


Fig. 4. Learning curve for the delta model. The solid curve is obtained from the best fit of the regression analysis on the theoretical learning curve.

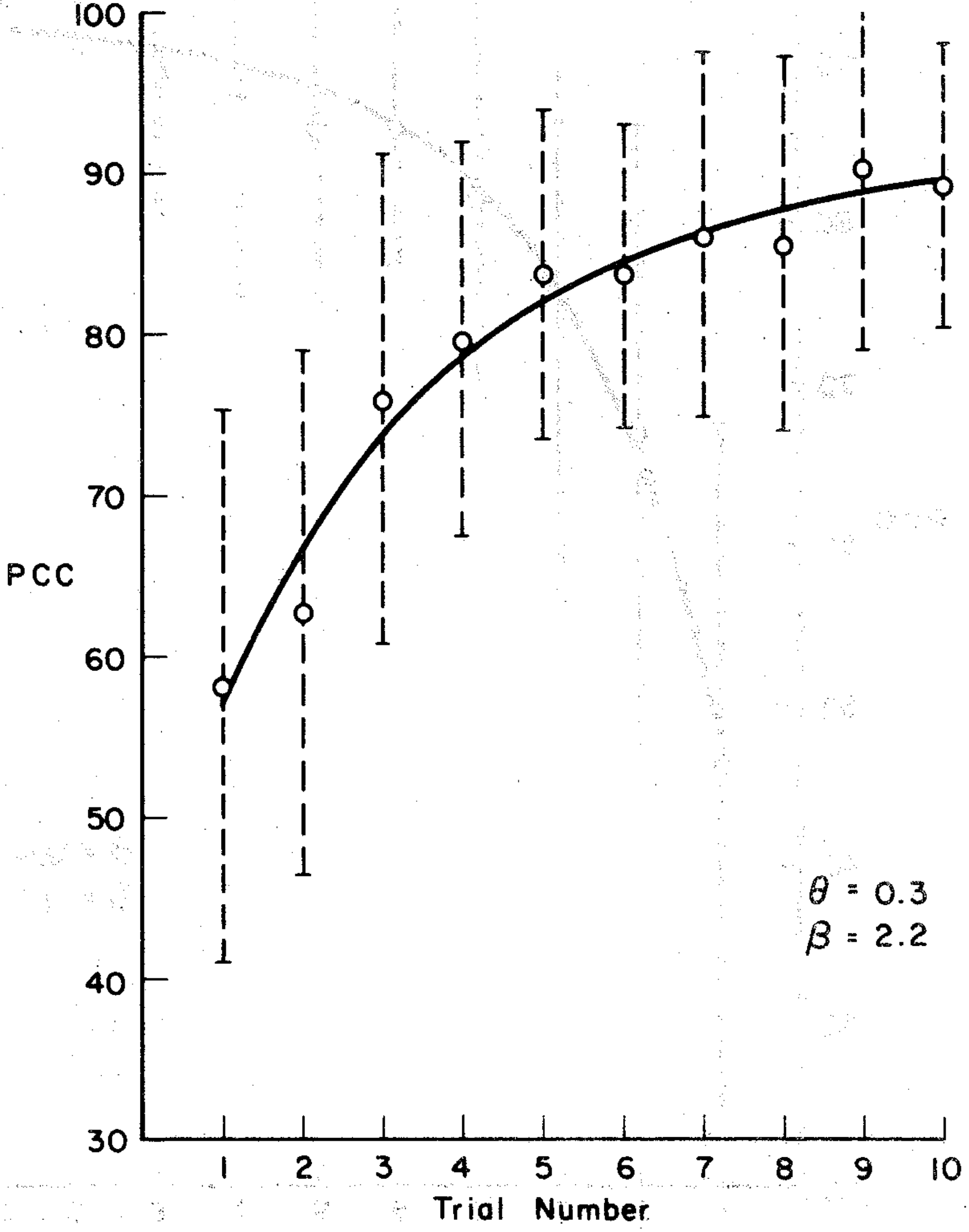


Fig. 5. Learning curve for the beta model. The solid curve is obtained from the best fit of the regression analysis on the theoretical learning curve.

of the curve in the conventional manner, using maximum likelihood estimators or other statistical techniques based on predictions of the particular learning model (Atkinson et.al.,1965). Instead we approach the problem of parameter estimation as a regression problem, with the mean learning curve of the form $Y = b_1 + b_2 Z$ where $b_1 = \pi$, $b_2 = p_0 - \pi$ and $Z = X^{n-1}$. Regression analysis for many values of X were performed on the learning curve data and the best fits, as determined by the R^2 and standard error of estimate statistics, were used to estimate the parameters (π, p_0) of the mean learning curve.

For the delta model the maximum R^2 statistic was .976 with an associated F value of 326 testing the statistical significance of the regression coeficients, and a standard error of estimation (s.e.e.) of .024. For X=.61 the parameter estimates π and p_0 were .923 and .477 respectively. Also, for X=.771, R^2=.937 and s.e.e=.039 with F value 119. Here the parameter estimates were π=.9994 and p_0=.53. Thus, the regression analysis of the delta model learning curve yields an asymptotic recognition rate above 92 percent with a 100 percent asymptotic recognition rate also giving a good fit to the learning curve.

For the beta model learning curve, similar analysis gave the greatest R^2=.969 with s.e.e.=.02 and F=247. Here X=.72, π=.914 and p_0=.572. The largest π was .94 with p_0=.592, X=.80, R^2=.80, s.e.e.= .025, and F=159. Again asymptotic recognition rates above 90 percent were found with the best fit at 91.4 percent and the maximum asymptotic rate of 94 percent with significant F values.

4.4 Parameter Grid Spaces

For simplicity in computation and discussion, we us the average of the ninth and tenth trials as the asymptotic approximation, although one cannot be certain that asymptote has been reached by the tenth trial. In our discussion we consider two different asymptotoic maxima, the group asymptotic maximum displayed in the learning curves and parameter grid spaces and the individual asymptotic maxima shown in the later figures. The group asymptotic maxima are obtained by averaging over the subject's individual asymptotic recognition rates for each grid point and selecting the maximum, while the individual maxima are simply the best asymptotic recognition rates for each child in his parameter space. The grid points for individual maxima may or may not coincide with the points for the group maximum. We use the group asymptotic value as our recognition rate, although the mean of the individual maxima is greater, in recognition of the importance of a single parameter setting generalizable across children.

To further illustrate the structure of the delta model, consider Table 4, which shows the percentage of asymptotic correct classifications averaged over the 14 subjects as a function of the parameters θ and δ. The parameter space displays a definite and regular structure for the delta model. The group maximum is 94.1 percent at grid point $\theta = .4$, $\delta = .1$.

Insert Table 4 about here

Similarly, we display in Table 5 the structure of the asymptotic percentage of correct classification over a grid of parameter settings

Table 4

Asymptotic Percentage of Correct Classifications,

Delta Model

	Delta										
Theta	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
1.00	78	78	78	78	78	78	78	78	78	78	68
.90	79	82	86	87	87	88	82	68	50	40	37
.80	80	88	89	89	89	85	77	62	47	41	36
.70	82	89	91	91	89	84	71	56	44	39	36
.60	83	91	92	91	88	82	69	53	43	35	36
.50	83	93	92	90	89	81	66	63	41	34	41
.40	85	94	92	91	88	79	65	49	39	37	36
.30	86	94	92	90	87	79	62	45	40	38	38
.20	87	92	92	90	85	74	56	44	42	39	40
.10	84	91	90	89	81	69	52	43	41	37	40
00	59	87	86	83	71	61	48	44	40	37	49

for the beta model. The parameter space for the beta model is noticeably flat even out to values of $\beta=5.0$. The group maximum is 89.8 percent at grid point (0.3,2.2).

Insert Table 5 about here

4.5 Theta Process

Again we consider the theta process alone. Figure 6 gives the error rate from column one of Table 4 ($\delta=0$). We note that to the accuracy of the curve the minimum occurs at the same value of theta that Table 1 predicts for the minimum of the expected distance of an utterance vector to its representation vector on the tenth trial. This striking correspondence between the minima of the error rate and the minima of the expected distance lends strength to analysis in terms of distances. Note that this analysis holds for two dissimilar situations, experiment 1 with a 2-word vocabulary and experiment 2 with a 14-word vocabulary.

Insert Figure 6 about here

4.6 Individual Asymptotic Maxima

So far we have been considering the group asymptotic maximum using one grid point for all subjects. This is important from an operational point of view, since when dealing with many children in a CAI curriculum it would be useful to have a general parameter setting good for all students. In Figures 7 and 8 and Tables 6, 7, and 8 we examine distributions of individual maxima over the parameter space in a further com-

Table 5

Asymptotic Percentage of Correct Classifications,

Beta Model

	Beta												
Theta	1.0	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0	2.1	2.2
1.00	78	78	78	79	79	79	79	79	79	80	78	78	78
.90	80	79	80	80	80	80	81	81	82	82	81	81	80
.80	80	81	81	82	81	81	82	82	82	82	82	82	83
.70	82	82	82	83	83	82	83	83	84	83	84	83	83
.60	83	83	84	83	83	85	84	84	84	83	84	84	84
.50	84	84	84	85	86	85	86	86	86	87	87	87	87
.40	85	86	86	86	87	88	87	88	88	88	89	88	88
.30	86	87	87	87	88	87	88	88	89	89	89	89	90
.20	87	88	88	89	88	88	88	88	88	88	88	88	88
.10	84	84	84	85	86	86	87	87	87	87	87	87	87
.00	58	62	63	64	67	67	68	68	68	70	69	69	71

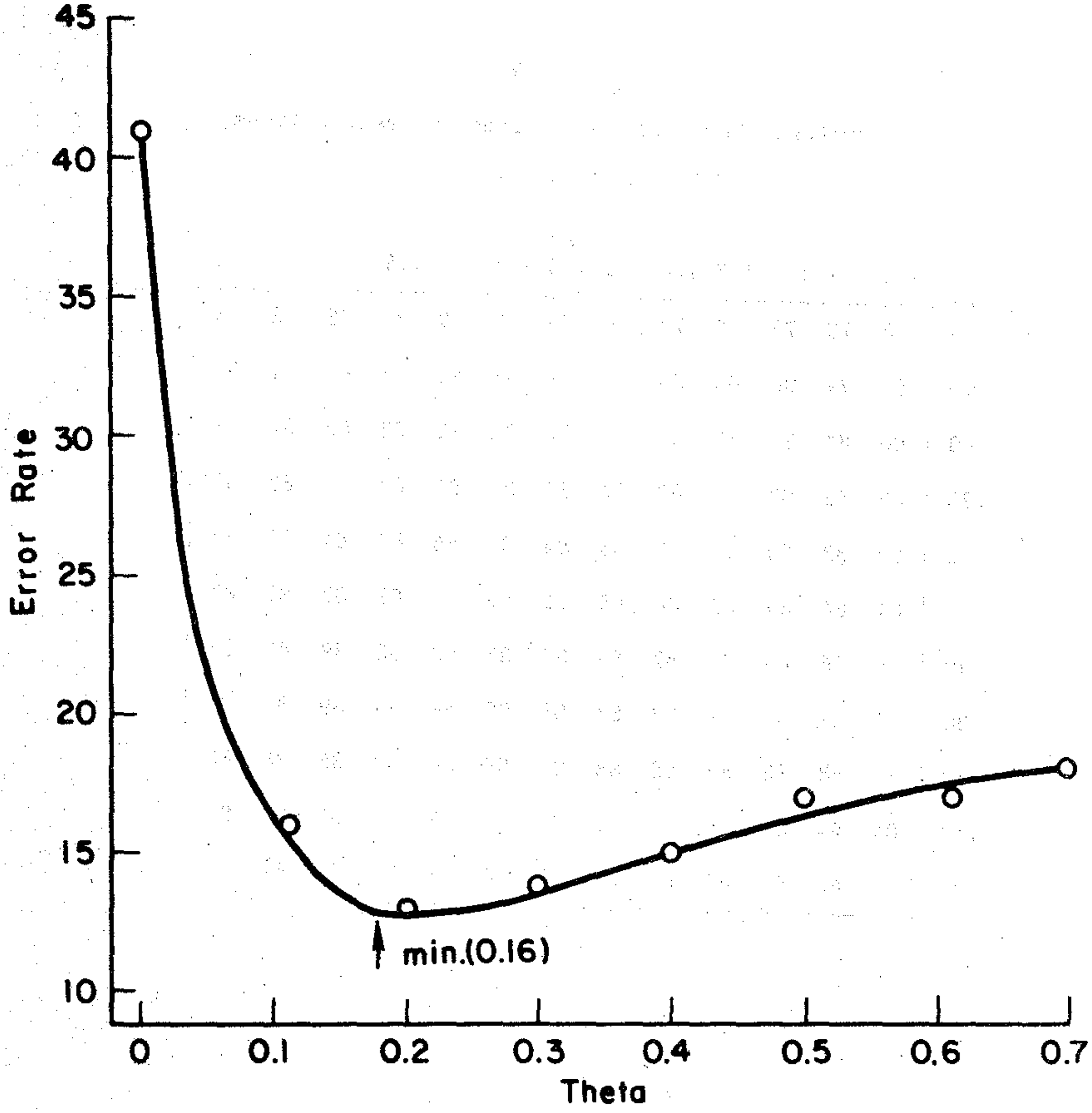


Fig. 6. Error-rate curve for the theta process alone. Note the agreement of the minimum with the prediction of Table 1 for the minimum of the expected distance of an utterance vector to its representation vector.

parison of the beta and delta models. We see in Figure 7 a 3 percent overall improvement for the beta model with individual improvements of as much as 13 percent for one subject when individual maxima are used instead of the group maximum. For the delta model the improvement was only 1.6 percent with the largest individual improvement being 3.6 percent. As can be seen from Tables 4 and 5 the delta model displays more regularity of structure about its group maximum than the beta model does.

Insert Figure 7 about here

4.7 Comparison of Individual Maxima for Beta and Delta Models

Note in Figure 8 the delta model does as well or better than the beta model in every case but one in this comparison. If we compared the individual asymptotes at the group maximum the delta superiority would be even greater. Hence, from these data from 14 children, we conclude that the delta model produces better recognition than the beta model.

Insert Figure 8 about here

4.8 Distribution of Optimal Parameter Settings

The distribution of optimal settings in the parameter space for the two models is displayed in Tables 6 and 7. For the delta model the asymptotic PCC for the grid point $\theta=.4$, $\delta=.1$ is consistently close to the individual maximum value for all subjects. Nine subjects attained

---

Insert Table 6 about here

```
                       Delta Model

6
5                                                              X
4                                           X        Y         X
3                                           X        Y         X
2                                           Y        X         X
1               X                           X        X         X
0----------.------------.------------.------------.------------.
           80           85           90           95          100
                            PCC ->           mean=95.7
```

```
                       Beta Model

5                                                       Y
4                                                       Y
3                                                       X      X
2                                         X      X      Y      X
1            X      X                     X      X      Y      X
0--.---------.---------.---------.---------.---------.---------.
  70        75        80        85        90        95        100
                           PCC ->      mean=92.6
```

X=male subject
Y=female subject

Fig. 7. Distribution of individual maximum asymptotic percentage of correct classification for the Delta and Beta models.

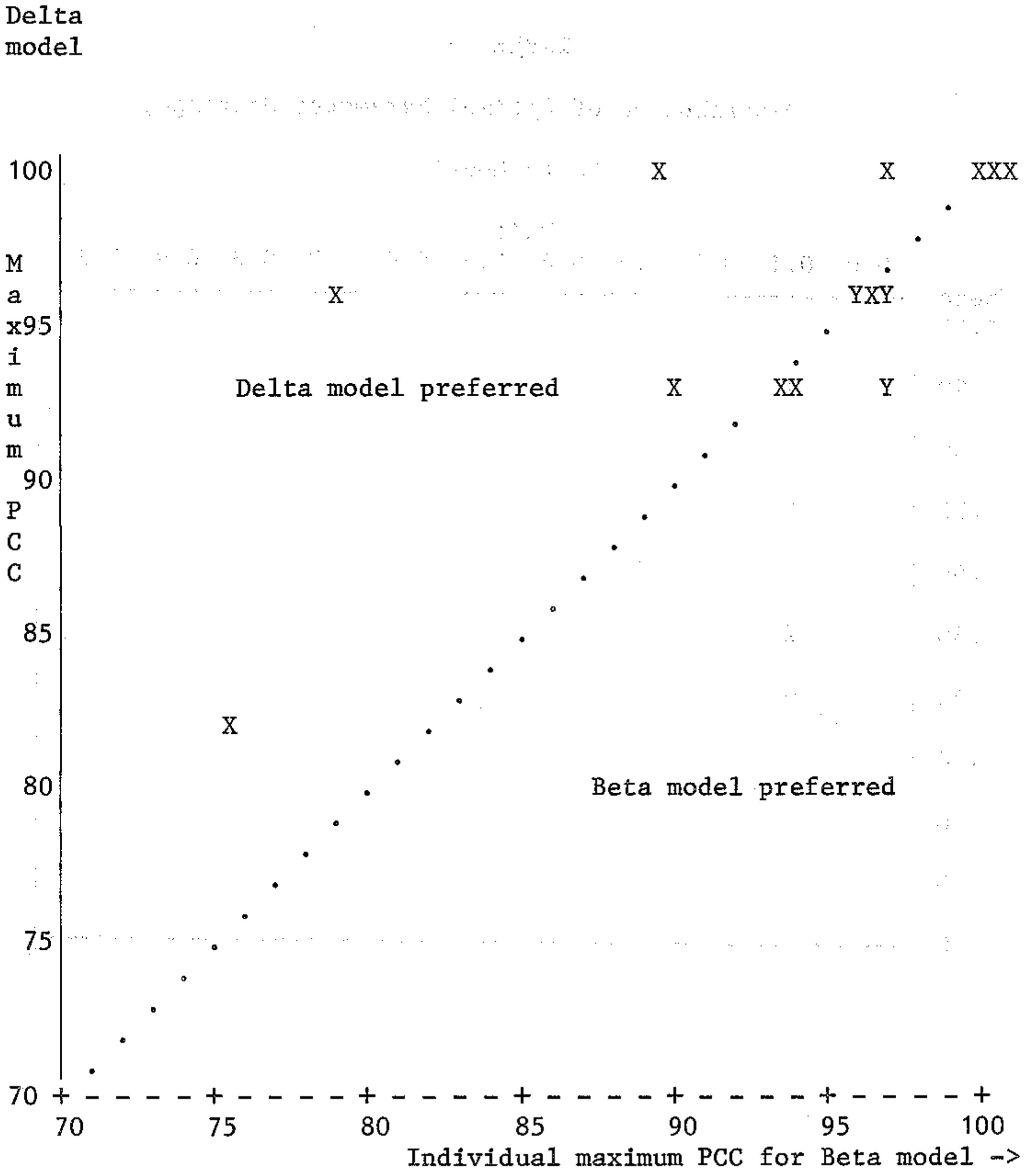


Fig. 8. Comparison of individual maximum percentage correct for the Beta and Delta models.

TABLE 6

Distribution of Optimal Parameter Settings,

Delta Model

					Delta						
Theta	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
1.00											
.90		1									
.80											
.70		1									
.60				1							
.50		2									
.40		9									
.30											
.20	1										
.10											
.00											

individual asymptotic maxima at this grid point, the group asymptotic maximum. For the five subjects who had different individual asymptotic maxima, the difference between their maximum recognition rate and their recognition rate for the grid point $\theta=.4$, $\delta=.1$ is only 3.6 percent for each subject. The beta model (Table 7) again shows more range and less definite structure than the delta model with 8 of 14 subjects having individual maxima distinct from the group maximum.

Insert Table 7 about here

4.9 Age Dependancy of Recognition Rate

From the results in Table 8 we can determine almost no age dependence for the recognition rates of children in the age range of 6 to 13 years old.

Insert Table 8 about here

5. EXPERIMENT 2, PART B

5.1 Description

The follow-up experiment was designed to determine whether or not the results of 2A had valid correspondence to actual working conditions of real-time recognition in a child's learning situation. It entailed investigating recognition rates for a telephone CAI mathematics curriculum with audio output based on Dial-A-Drill (Computer Curriculum Corp., 1971), which incorporates the delta model into the learning scheme. The system presently runs in a fully automatic mode in that a telephone call

TABLE 7

Distribution of Optimal Parameters Settings,

Beta Model

	Beta										
Theta	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0	2.1	2.2
1.00											
.90											
.80											
.70											
.60											
.50											
.40									1		1
.30								1			6
.20									1	1	1
.10						1					1
.00											

TABLE 8

Distribution of Individual Maximum Asymptotic PCC

as a Function of Age

Delta Model

```
                         Age of Subjects
       6        7       8       9       10      11      12      13
PCC----------------------------------------------------------------
100|                    X               X       XX               X|
  8|
  6|    Y                                  XY                    X|
  4|
   |            X            X                                   X|
  2|                                 Y
 90|
  8|
  6|
  4|
  2|            X
 80|-------------------------------------------------------------|
```

Beta Model

```
                         Age of Subjects
        6       7       8       9       10      11      12      13
PCC----------------------------------------------------------------
100|                                    X       X               X |
  8|                                                              |
  6|    Y                          Y      XY    X                 |
  4|                                                              |
   |           X           X                                      |
  2|                                                              |
 90|                                                              |
   |                    X                                        X|
  8|                                                              |
  6|                                                              |
  4|                                                              |
  2|                                                              |
 80|                                                              |
   |                                                             X|
  8|                                                              |
  6|                                                              |
   |            X                                                 |
  4|                                                              |
  2|                                                              |
 70|--------------------------------------------------------------|
```

placed to (415) 321-TALK is automatically answered by a coupler that connects the outside phone line to an Institute high-speed data line (HSL10). The TENEX executive program of the PDP-10 has been modified so that upon sensing the input data over HSL10 it will log in a job and run a prespecified program.

The mathematics curriculum is extremely simple. It uses only the 14-word vocabulary described in the previous experiment, and randomly generates five types of questions (sum, difference, product, integer division, and ratios). The computer-student dialogue consisting of these arithmetic questions, produces audio output that is accessed quickly from magnetic disk and transmitted to the student over the telephone.

5.2 Dial-A-Drill Curriculum Structure

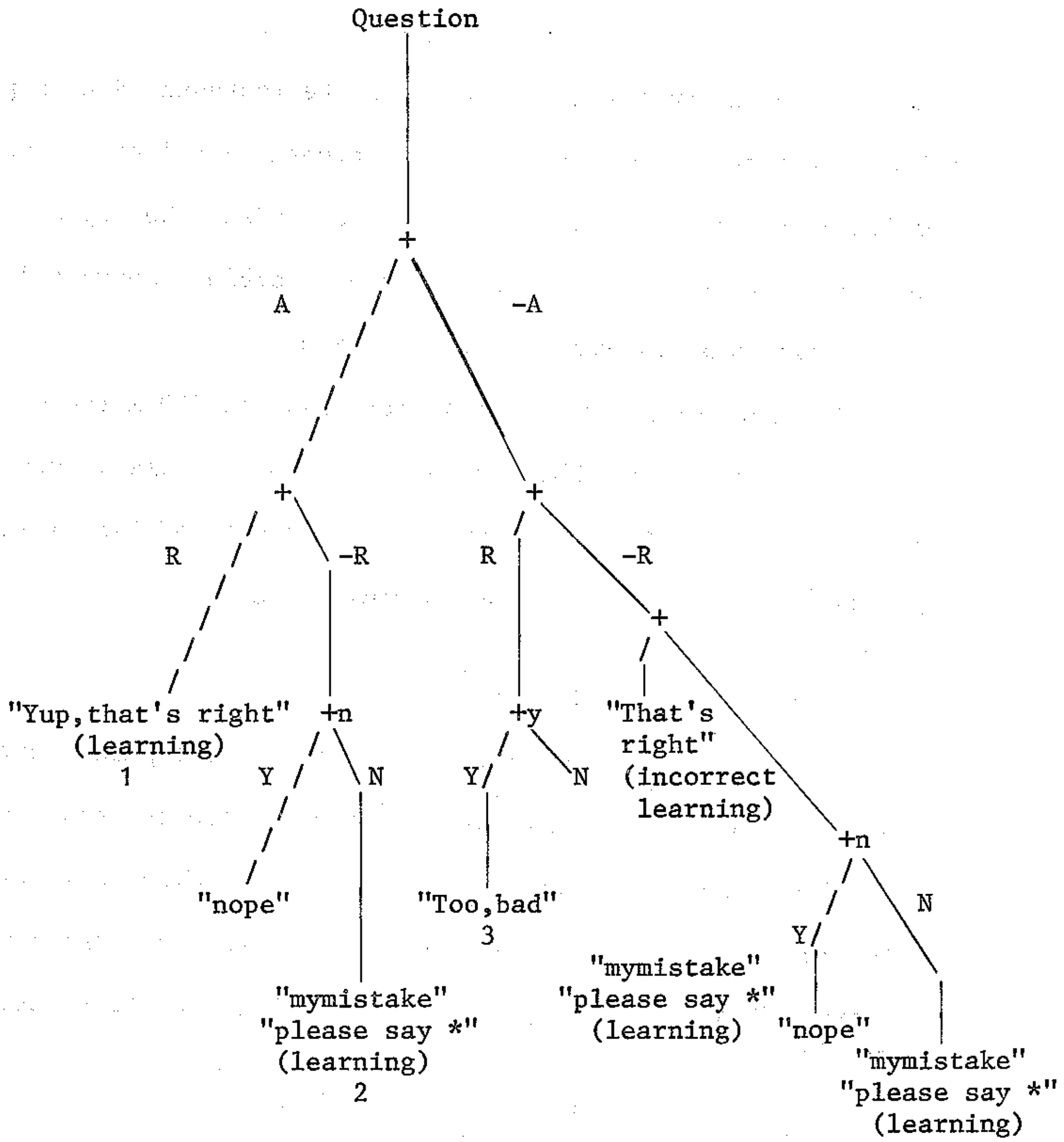
The curriculum structure is indicated in Figure 9. A typical segment of the dialogue could consist of the following:

Insert Figure 9 about here

```
COMPUTER: WHAT IS 7 MINUS 3 ?
STUDENT:  4
COMPUTER: YUP, THAT'S RIGHT.                 Path ending at  1
                                              (see tree)

COMPUTER: 6 OVER 3 IS WHAT ?
STUDENT:  2
COMPUTER: DID YOU SAY 0 ?
STUDENT:  no
COMPUTER: SORRY, MY MISTAKE. PLEASE SAY 2.
STUDENT:  2
COMPUTER: THANK YOU!                         Path ending at  2

COMPUTER: 5 TIMES 0 IS WHAT ?
STUDENT:  5
COMPUTER: DID YOU SAY 5 ?
STUDENT:  yes
COMPUTER: TOO BAD, 5 TIMES 0 IS 0.           Path ending at  3
```



SYMBOLS

A	Correct answer given by child
-A	Incorrect answer given by child
R	Recognized by computer
-R	Misrecognized by computer
+	Node
+n	Node with "no" response always
+y	Node with "yes" response always
Y	Computer thinks yes was said
N	Computer thinks no was said
*	Correct answer (to be repeated by student)

Fig. 9. Tree diagram for learning in the mathematics curriculum.

Each of the above three dialogues can be represented as a path along the learning tree shown above. We use noncontingent learning for the delta model on all correct responses and also update the representation vector on all requested repetitions. The one possible incorrect learning node on the tree was not realized in practice.

Seven subjects from Part A each answered 100 mathematics exercises from the curriculum. The recognition mechanism was loaded with a state vector for each subject obtained from the data of Experiment 2A using the delta model at the optimal parameter settings.

5.3 Comparison of Parts A and B

The resulting recognition rates for the telephone curriculum are shown in Figure 10 and average 13 percent below the best recognition rates for the subjects in Experiment A. The decrease in recognition rates in Part B can be accounted for by educational and psychological factors. We did not have an introductory session to acquaint the child with the system. Also, in an effort to approximate natural home conditions we gave no instructions to the child about speaking carefully. When faced with a mathematical question instead of a mere request to repeat a number the student sometimes stammered or changed his mind in the midst of an utterance (e.g., "ONE--NO-TWO!"), which had obvious degrading effects on the recognition rate. Observe that even under these conditions the recognition rates are all above 75 percent.

Insert Figure 10 about here

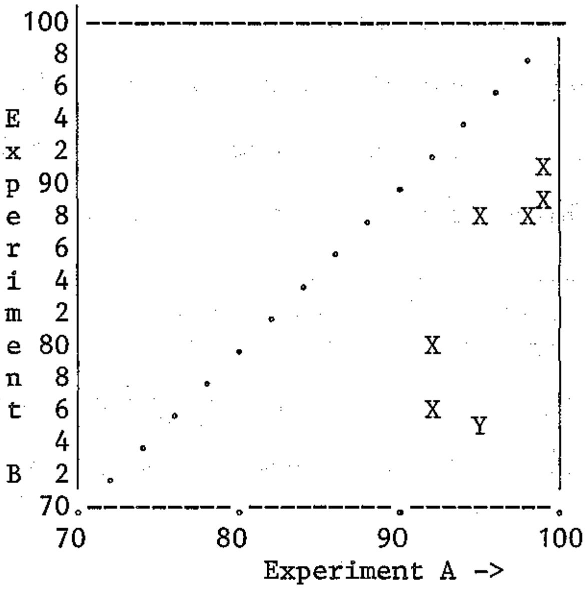


Fig. 10. Comparison of recognition rates on experiments 2A and 2B for the seven subjects completing both experiments.

5.4 Confusion Matrix for Experiment 2B

In Table 9 we present a confusion matrix of the numbers 0-9 for the seven subjects in experiment 2B. Each element of the confusion matrix (c_{ij}) represents the number of events where the utterance was i and the classification was j. Thus the matrix entry $c_{2,0}$ indicates the number of events where 2 was said and the computer misclassified the utterance as 0.

Insert Table 9 about here
---

6. SUMMARY AND CONCLUSIONS

We have constructed and tested two models of learning processes for the purpose of computer recognition of human speech over the telephone. The delta model was found superior to the beta model in all comparisons. For the delta model a regression analysis on the learning curve yielded a 92.3 percent recognition rate for 14 subjects ranging in age from 6 to 13 years old. When the individual approximate asymptotic maxima are used the recognition rate climbs to 95.7 percent. All the recognition was done using a standard home telephone.

It should again be emphasized that we are conducting real-time recognition in a time-sharing environment without any linguistic restrictions and with relative computational simplicity. Consequently, the system can be used on any language from Swahili to English.

We also tested the recognition system on an elementary mathematics curriculum conducted entirely over the telephone in Experiment 2B. From our observations we found that the children seemed quite tolerant of

TABLE 9

Confusion Matrix for Experiment 2B

	0	1	2	3	4	5	6	7	8	9
0:	159	2	8	0	1	0	2	0	0	0
1:	3	79	2	0	8	2	0	1	0	5
2:	12	0	64	1	0	0	2	1	0	0
3:	2	1	1	45	0	0	1	1	3	0
4:	3	2	1	0	54	0	1	0	0	0
5:	0	7	1	0	1	35	0	1	0	8
6:	2	0	3	0	0	0	42	0	4	0
7:	2	1	2	0	0	0	1	38	0	0
8:	4	0	1	3	0	0	3	0	40	0
9:	0	0	0	2	0	0	0	2	0	27

nonperfect recognition and, indeed, were amused when the computer made a mistake.

In our efforts we are approaching speech recognition from the direction of machine learning. In analyzing experiment 2A we see that learning indeed occurs and is amenable to theoretical analysis for the purpose of predicting the learning performance from the structure of the model. Future efforts will be directed toward deriving the exact form of this performance and toward making deeper comparisons with human learning theory.

APPENDIX

Previous Speech-Recognition Work at IMSSS

Camille Bellissant
Stanford University

1. INTRODUCTION

The aim of this work was to run some preliminary experiments using audio for both input and output in a computer-assisted instruction (CAI) program.

The output part, i.e., speech production, was handled by an existing program that gives good results for short sentences. The production is not done by synthesizing but by digitizing spoken words which are later concatenated to produce a sentence. The random access to the digitized records on the disk allows quick retrieval and, when the computer is not overloaded, permits a continuous audio output.

For the input part, i.e., speech recognition, it was decided to begin by adopting the system developed by Raj Reddy and Pierre Vicens at the Artificial Intelligence Project, Stanford University (Vicens, 1970). This choice was justified by the effectiveness of the system for recognizing isolated words belonging to a small vocabulary (about 50 words, which is large enough size for our purpose).

In the next section we describe the Vicen's program and our modifications of it.

2. THE MODIFIED VICEN'S PROGRAM

In his thesis, Vicens (Vicens, 1969) presents the techniques and methodology he used in building the system.

2.1 Preprocessing.

The audio message to be recognized is first preprocessed by hardware. Three filters (150-900 Hz, 900-2200 Hz, 2200-5000 Hz), corresponding roughly to the first there formants of voice, and an analog to digital converter produce for each frequency band and for each sample of 10 ms of sound the maximum amplitude (peak to peak) and the number of zero-crossings of the amplitude-time function. The data are transmitted through a high-speed line to the software preprocessor, which normalizes the amplitudes.

2.2 Segmentation.

After the hardware and software preprocessing, the data are treated by the segmentation procedure. This consists of grouping the minimal segments of 10 ms into wider segments presenting roughly the same acoustic characteristics (sustained segments) and isolating the others into transitional segments.

Although some errors can occur in this grouping, and a secondary segmentation procedure corrects the possible errors by looking at the variation of parameters in the sustained segments and at the local maxima and minima of the amplitude parameters in the transitional segments. If the variation of parameters in a sustained segment exceeds a certain limit, or if a transitional segment presents a local extremum, the segment is divided into smaller ones.

The last part of the segmentation is the combining process whose purpose is to group together acoustically similar secondary segments. The sustained segments are extended onto the transitional segments if the parameters are too different.

2.3 Classification.

After the segmentation process, most of the transitional segments, which do not contain pertinent information, are eliminated. The purpose of classification is to assign linguistic labels to the sustained segments. The phoneme groups are fricative, vowel, stop, consonant, nasal, and burst. The vowels are subclassified into nine categories with respect to their zero-crossing parameters. The discrimination into phoneme groups is accomplished by comparing the amplitude and zero-crossing parameters of the segments with known values in acoustic phonetics.

The results of the previous processes are summarized in an internal representation of the speech utterance that is used for all the storing, retrieving and matching processes.

2.4 Recognition.

The recognition of words is accomplished by retrieval of previously learned messages. This learning consists of reducing the internal representation of the speech utterance and storing the reduced form in a dictionary. The size of such a reduced record is about 1000 bits of continous storage for an average sound of 1 second.

The dictionary is provided with two independent list structures depending on the phonetic representation fo the message (number of vowels and unvoiced fricatives), and the print name of the message. During the recognition phase, the dictionary organization allows a quick candidate list to be constructed in the following stages:

1. Elimination of all candidates whose relative positions of vowels and fricatives are different from those of the incoming message.

2. Elimination of all the candidates with strictly different vowel zero-crossing characteristics.

3. Elimination of all the candidates having low-vowel similarity scores obtained by comparison with the incoming message.

The first elimination is obtained directly from the dictionary, which holds the relative position of vowels and fricatives for each recorded utterance. The second elimination is accomplished by using a table that defines crude dissimilarity values between each pair of vowels on the basis of their earlier classification into subcategories. At this stage the list of candidates is reordered, so that the most similar candidates are placed first.

The third elimination is done by computing a similarity between the incoming message and all the entries in the candidate list. First, a segment synchronization procedure is called to create linkages between the segments of the two representations. The similarity values obtained for each pair of linked segments are stored for the selection process that chooses the candidate with the higher similarity coefficient. If one of the candidates reaches a score greater than or equal to 95 percent,the selection process immediately stops and returns the candidate print name. Otherwise, each time a good similarity score is obtained (>80 percent) the candidate list is rearranged in order to place first all entries having the same print name as that of the present candidate. When the list is exhausted, the candidate with the best score is chosen if the similarity score is at least greater than 75 percent. If none of the candidates presents such a score, the selection process is reinitiated with a new list of candidates having small differences in the phonetic representation (number and relative position of vowels and unvoiced fricatives). If no candidate can be found in the dictionary, it means that the incoming message cannot be recognized, and the user is invited to enter its print name. At this time, the dictionary is augmented by the representation of the new message, which can be used afterwards as a possible candidate for a further utterance.

2.5 Modifications.

The original Vicens' program as described above was written in FORTRAN for the PDP-10 at the Stanford Artificial Intelligence (AI) Project. The program was rewritten for the PDP-10 at the Institute in SAIL, which is a high-level language that is a superset of ALGOL and that has been developed at the AI project.

Second, hardware that has performance characteristics very similar to the hardware on the AI PDP-10 was designed and constructed by Ron Wizelman of the Institute staff. There is a slight difference in the handling of the incoming data as given by the hardware. The Vicens' program was working within a "spacewar" environment in order to impose priority over other users while listening to the sound. We use a high-speed line that gives good results at all times for a continuous input. In order to allow the user not to speak as soon as the program is ready to listen to him, we have implemented two thresholds. One is hardware, the other is software. The first system is a simple potentiometer that inhibits the hardware equipment as long as the amplitudes are under a certain value. This value is adjustable and can eventually become zero. As soon as this threshold value is exceeded, the hardware begins to transmit data to the program and keeps transmitting even after the amplitudes again drop under the threshold value. This delay, which is also adjustable, is necessary to allow small silences in the utterance without interrupting the transmission. Our first experiments with this hardware threshold have shown some loss of data in the very beginning of each utterance, due to the positive value-fixed threshold.

In order to avoid this loss of data, we experimented with kicking the microphone that started the hardware and speaking just after the kick. The effect of the kick has been eliminated by software and so no data were lost. Besides the inelegance of such a method, we found it difficult to apply to all kinds of microphones, especially telephones. So we introduced the following process. The hardware threshold is set to zero, so the hardware is always ready to transmit data. The software procedure reads only three samples of sound (0.03 sec) and computes the averaging amplitudes and zero-crossings. If these values are under a threshold, three new samples are processed, and so on. If the values are above the threshold, the procedure fills up the input buffer (1.5 sec). The 'tail' of the utterance, i.e., samples with low amplitudes at the end of the message, is then eliminated so that only the relevant values are subsequently processed by the segmentation procedure. When the computer is overloaded, this method (the 'software kick') sometimes produces a loss of one sample (10 ms), which is actually the smallest amount of data that can be lost.

The other differences we introduced in the Vicens' program concern the selection process. First, the value of the similarity threshold (95 in the original program) which is used when one examines the candidate list, was changed by an interactive command. We are concerned with the best choice of the threshold value for different sets of words. Intuitively, the larger the value, the more demanding the system when it tries to accept a candidate as a proper answer. Sets of the words with large phonetic dissimilarity can be processed with a low threshold and a consequent saving of time.

The second difference is related to the use of the system in pedagogical experiments in elementary arithmetic. The purpose of these experiments is to ask the user the results of operations on numbers. In this case, for each question there exists one and only one possible answer. When the answer is incorrect, we do not try to recognize the specific value that was uttered. For example, after the question "how much is three plus four?" we are only interested in the comparison between the uttered answer and 7. If it is not 7 we do not try to know whether it was 6, 8, or something else. In this situation, the recognition process can be considerably accelerated by limiting the candidate list to those that have the same print name as the expected answer. We found that in this way the answer processing is faster than the time spent to utter it, which offers some hope for communication by telephone when the nature of the messages to be recognized is well adapted to such a discrimination.

REFERENCES

Anderson, T. W. An introduction to multivariate statistical analysis. New York: Wiley, 1958.

Atkinson, R. C., Bower, G., & Crothers, E. An introduction to mathematical learning theory. New York: Wiley, 1965.

Bush. R. R., & Mosteller, F. Stochastic models for learning. New York: Wiley, 1955.

Computer Curriculum Corp. Dial-A-Drill. Mt. View, Calif.: Computer Curriculum, Corp., 1971.

Estes, W. K., & Suppes, P. (Eds.) Studies in mathematical learning theory. Stanford, Calif.: Stanford University Press, 1959.

Lamperti, J., & Suppes, P. Some asymptotic properties of Luce's Beta learning model. Psychometrika, 1960, 25, 233-241.

Vicens, P. Preprocessing for speech analysis. Project Memo No. AI-71. Stanford Calif.: Artificial Intelligence Laboratory, Stanford University, 1970.

Vicens, P. Aspects of Speech Recognition by Computer. Unpublished doctoral dissertation, Stanford University, 1969.