STATISTICAL ASSESSMENT OF SPEECH SYSTEM PERFORMANCE

STEPHEN L. MOSHIER*

DIALOG SYSTEMS, INC.
BELMONT, MASSACHUSETTS

Introduction

Since many of the participants at this conference have the task of evaluating and comparing different speech recognition systems that have been tested under various conditions, the author has collected in Part I some useful statistical rules of thumb which can be employed to normalize disparate experimental test results. Several types of elementary statistical analysis are illustrated; the reader is encouraged to continue in this spirit to analyze other cases. The rules are not widely known, but seem to have good predictive power. All of the ones presented here are accompanied by supporting empirical evidence.

Part II, the advertising part, describes some of the accomplishments and planned development activity of Dialog Systems, Inc. in accordance with the Workshop specification. Dialog's sole business is speech recognition. The company has successful operational field experience with its first major product, a multi-channel talker-independent system for verbal inquiries via ordinary switched network telephone input. Dialog presently has 45 employees, including a competent support and field service staff.

I. Methods for Normalization of Performance Test Results

Contrary to some beliefs, speech recognition systems obey the laws of nature. The small number of known quantitative rules are statistical in character, and they relate such variables as average recognition accuracy, vocabulary size, reject rate, false alarm rate, and the sizes of experimental training and test sets. The relations to be presented here seem to have good predictive power, and the author uses the illustrated analyses on a day-to-day basis to evaluate and compare different experimental results. It is very obvious that some such probabilistic rules must apply, though there are questions of detail and refinement of the statistical models to be resolved. It will be possible in the future to tie together many other experimental variables, but to do this it will be necessary for investigators to include more detailed experimental data in their reports and to test much larger populations than they have been accustomed to using, on the average.

*Mr. Moshier's paper was presented by Mr. Robert Osborn.

The ultimate capability of practical speech recognition systems has not been determined. The speaker verification system developed by Doddington's group, for example, probably does better than a human could do; the computer does not get tired, and can be programmed to notice identifying characteristics that people pay no attention to. In the various kinds of speech recognition, performance is limited by such things as insufficient data bases, mistakes in computer programs, and adherence to wrong theories; we are certainly quite far from any limits set by thermodynamics or information theory.

When various published data are normalized by means of the statistical rules to be described, it emerges that there has been essentially no fundamental progress in isolated word speech recognition since the first good techniques appeared in 1969-1972. On the other hand, there has been a great deal of progress in making the fundamental principles work in field applications, as well as in other areas.

Vocabulary Size

Most reports on speech recognition give a figure for recognition rate and vocabulary size. The law of nature is that recognition rate decreases with increasing vocabulary size. It is quantified by a statistical rule of thumb as follows:

Given that the input speech is word (or phoneme, or sentence) X_i , suppose that the machine is characterized by the probability that it will correctly reject the possible wrong choices x_j , $j \neq i$:

$$\Pr\{\text{correctly reject } x_j \mid x_i\} = r_{ji}. \qquad (1)$$

In the interest of deriving a simple formula, make the following assumptions: a) r_{ji} is about the same for all pairs of vocabulary words, so that it can be replaced by a constant value r. (In practice this is usually true except for a small number of troublesome words having high confusion probability; but if the relative proportion of troublesome words is constant the formula remains true for an appropriate choice of r. Thus a less stringent assumption is sufficient, namely that the distribution of values r_{ji} is dependent of vocabulary size.) b) The various correct rejection probabilities (1) are all statistically independent. This assumption permits getting the probability of joint events by multiplying. As in assumption a), it can be replaced by less strict conditions, but then the development becomes more obscure. It is <u>not</u> true for some types of joint events encountered in continuous speech recognition (see below). Let there be n words in the vocabulary; under conditions a) and b) the probability of correctly rejecting all of the wrong choices x_j, $j \neq i$ is

$$\Pr\{\text{correctly reject } x_{j_1} \text{ and } x_{j_2} \text{ and } \ldots x_{j_{n-1}} \mid x_i\} = \prod_{j \neq i} r_{ij} = r^{n-1}. \qquad (2)$$

This is the correct recognition rate of the system. Curves for talker-dependent and talker-independent isolated word recognition are given in Figure 1. These curves were drawn in mid-1974; there does not seem to have been much change since then except that they are perhaps a little truer now than they were before.

False Alarms

A related statistical rule has to do with false alarm events in word spotting. In this task, the event of interest is the joint detection of several acoustic segments in the right sequence. The unconditional probability of a false alarm for any one of the segments is assumed to be small in a small time window and independent of time. The distribution of false alarms is therefore Poisson with some rate function λ. Given the acoustic event x_1, certain following events are more likely to occur than others, on the average. Thus it is not possible to get the probability of a joint event by multiplying the individual event probabilities. A joint false alarm can be modeled quite closely, however, as a sort of Markov chain. It is assumed that if the first target segment x_1 is detected, the unconditional Poisson rate function for subsequent detection of the second event must be multiplied by some value α.

The Poisson law implies that the unconditional probability of not detecting the ith event is $e^{-\lambda i}$. If the conditional detection probability depends on the immediately preceding event but no earlier ones, the probability of the first two events jointly is

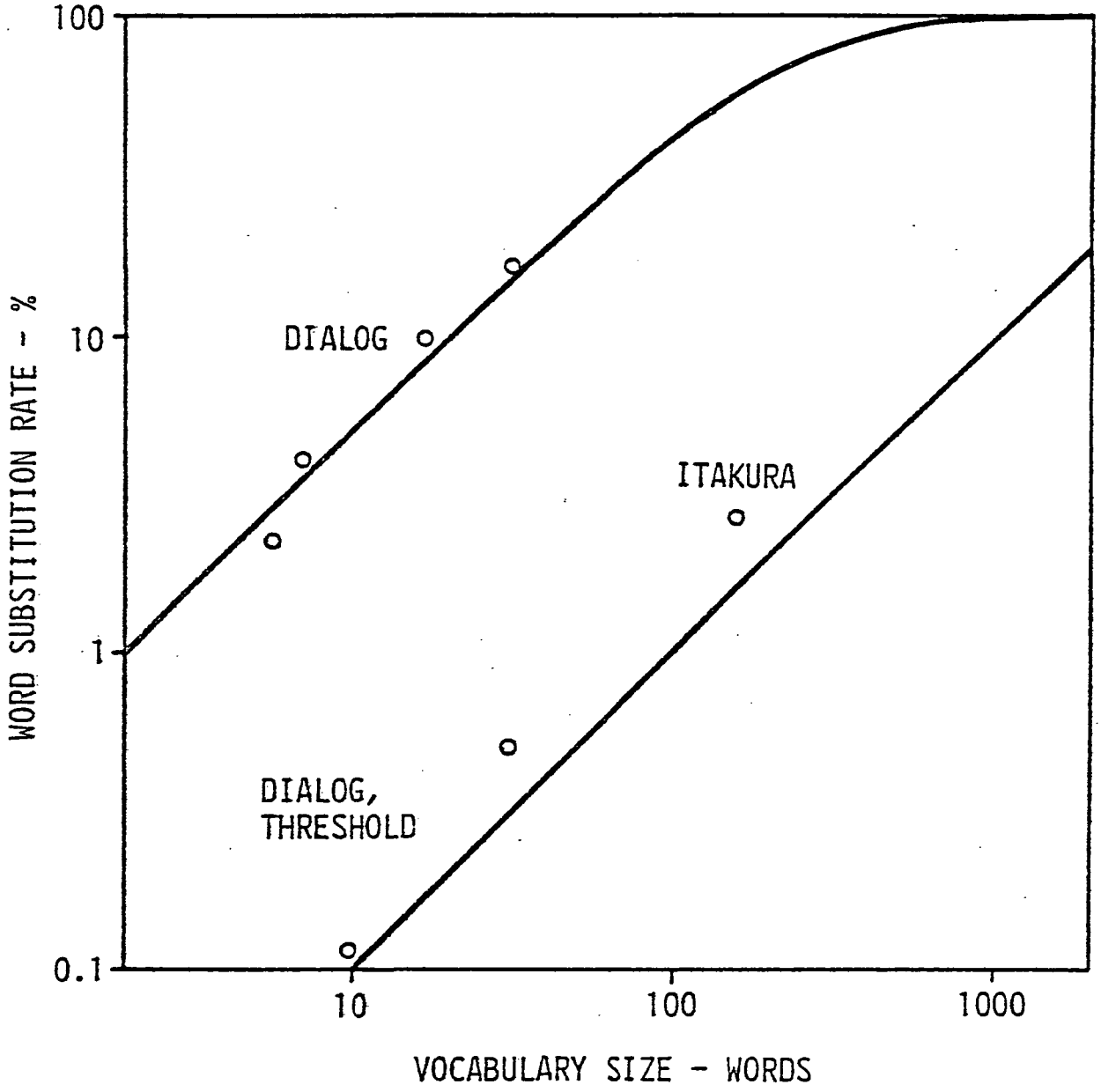
$$\Pr\{x_1 \text{ and } x_2\} = \Pr\{x_2 \mid x_1\} \Pr\ x_1 = (1-e^{-\alpha\lambda 2})\ (1-e^{-\lambda 1}).$$

The joint detection probability for n events is then

$$\Pr\{x_1 \text{ and } x_2 \text{ and} \ldots x_n\} = (1-e^{-\lambda 1})\ (1-e^{-\alpha\lambda 2})\ (1-e^{-\alpha\lambda 3})\ldots(1-e^{-\alpha\lambda n}). \quad (3)$$

Some experimental data are compared with this model in Figure 2. J.L. Baker, of IBM, has built a Markov model for correct detection events directly into a continuous speech recognition algorithm.

In addition to their ability to normalize the results of different experiments, these models to some extent permit one to separate and compare the statistics of language and the statistics of the recognition algorithm. The same models apply to isolated word and continuous speech recognition, except that an extra error contribution from the isolated word boundary detector is needed. There is no apparent reason why continuous speech recognition systems with performance as good or better than isolated word systems in comparable tasks should not be possible. Developmental results approaching the best isolated word techniques have already been reported.



Bottom curve: system tuned to individual talker.

Top curve: talker-independent performance for telephone speech.

Figure 1. Current State-of-the-Art Performance in Isolated Word Recognition

Event:	Joint false alarm rate predicted by equation (3) with $\alpha=2.4$:	Observed joint false alarm rate:
$\{A_1, A_2\}$	24.7	27
$\{A_3, A_4\}$	80.8	67
$\{A_1, A_2, A_3\}$	3.5	4
$\{A_1, A_2, A_3, A_4\}$	1.5	4

Figure 2. Comparisons of observed false alarm rates with predictions made by equation (3) for multiple pattern phrases in continuous telephone speech. Observed unconditional probabilities $Pr\{A_i\}$ of false detection in an interval ≈ 0.17 second are: $Pr\{A_1\} = 0.051$, $Pr\{A_2\} = 0.110$, $Pr\{A_3\} = 0.084$, and $Pr\{A_4\} = 0.205$

ROC Curves

Normalization of different test results frequently requires an estimate of the relation between reject (no decision) rate and correct recognition rate. Relatively few reports give this function, so there is little published information to use as a check on the model. Hence, the following model is not known to apply to anything but the Dialog system. The algorithm produces a goodness of fit score for each decision of interest. Over many trials the fit to a particular reference template has a distribution which is not Gaussian; but the difference between the scores for the closest fitting template and the correct choice template does seem to be approximately Gaussian. A reject criterian based on this function is illustrated in Figure 3. The model yields a family of parallel lines on a probability scale graph; thus only one measurement is required to determine which line corresponds to the system under test. A similar model can be derived for the more commonly used reject criterion in which the input is rejected if no template matches it sufficiently well.

Probable Error of Measurements

The vast majority of published reports in the field do not contain enough information to establish error estimates for the claimed numerical test results. From this symptom and many others which vary from paper to paper, one may justifiably conclude that the average experimental quality of current speech R&D investigations is absolutely terrible.

To obtain statistical confidence intervals for a parameter, it is necessary to known something about its probability distribution. Most reports contain no helpful information whatever, so one can only guess. For small sized test samples a non-parametric approach can be taken: the experiment is modeled as a series of Bernoulli trials with a binomial distribution of test scores. This method produces seemingly pessimistic estimates of the probable range of random sample test results; but caution and pessimism are the correct attitudes to adopt when interpreting small-sample statistics. Tables of confidence intervals for the binomial distribution are available in an RADC report.

For medium sample sizes (more than 30 trials per talker and more than 30 talkers) a better procedure is to assume that the total number of errors has a Poisson distribution. There is some evidence that the Poisson law is actually a good model for pattern recognition methods which show a low error rate; but the main advantage is that the Poisson distribution has only one parameter, so the answer can be looked up immediately in Figure 4. To use the table, count up the total number n of errors observed in the experiment and find the upper and lower bounds of the desired confidence interval from the appropriate columns. The total number of trials in the experiment is immaterial; the tabulated figures represent total numbers of errors, and must be divided by the

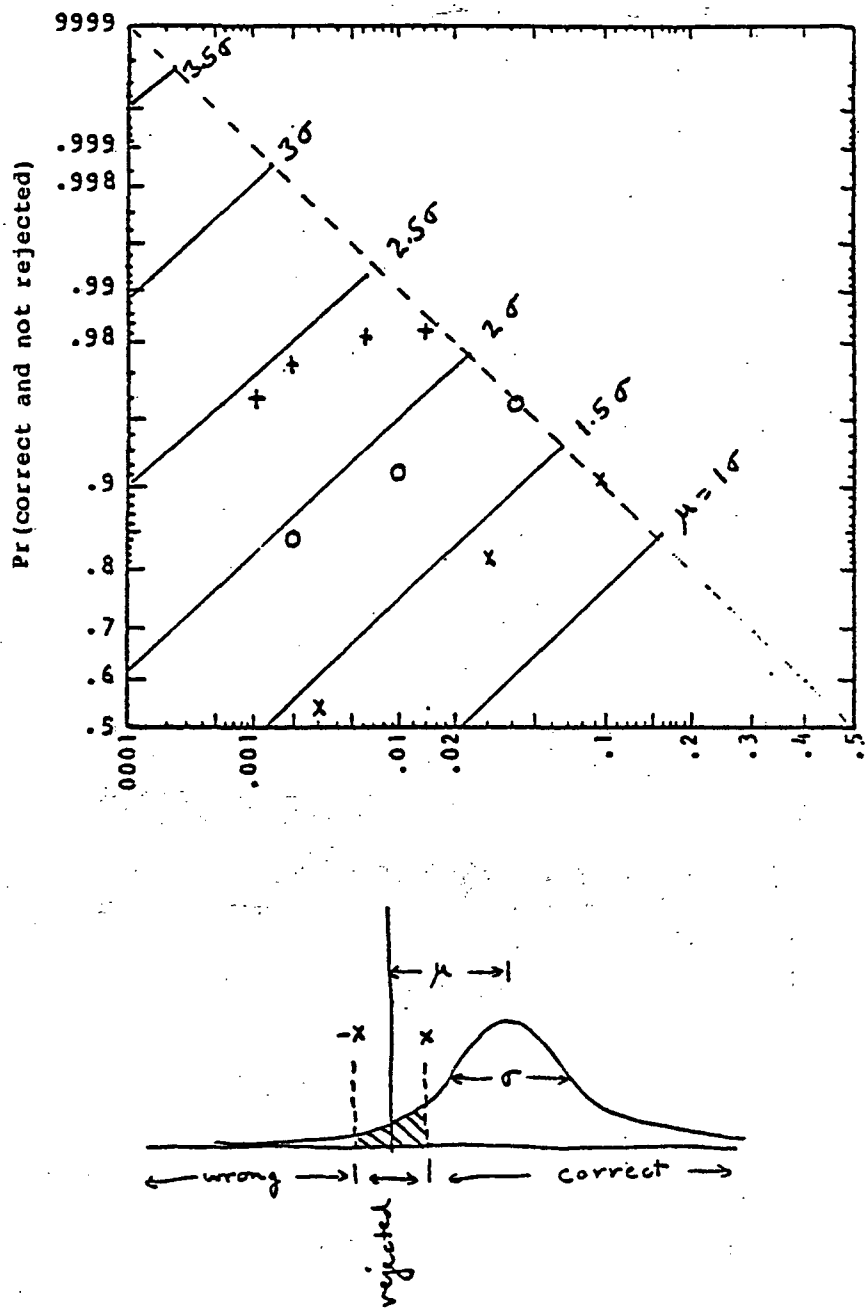


Figure 3. ROC curves for a quasi forced-choice decision rule. Experimental data are plotted for a selected 8-word vocabulary (+), the 10 digits (0), and a 34-word vocabulary (X) in a discrete word recognition task with telephone speech and many talkers.

n	c = .01	.05	.10	.90	.95	.99
0	0.010	0.051	0.105	2.303	2.996	4.605
1	0.149	0.355	0.532	3.890	4.744	6.638
2	0.436	0.818	1.102	5.322	6.296	8.406
3	0.823	1.366	1.745	6.681	7.754	10.045
4	1.279	1.970	2.432	7.994	9.154	11.605
5	1.785	2.613	3.152	9.274	10.513	13.108
6	2.330	3.285	3.895	10.532	11.843	14.571
7	2.906	3.981	4.656	11.771	13.148	16.000
8	3.508	4.695	5.432	12.995	14.435	17.403
9	4.130	5.425	6.221	14.206	15.705	18.783
10	4.771	6.169	7.021	15.407	16.962	20.145
11	5.428	6.924	7.829	16.598	18.207	21.490
12	6.099	7.689	8.646	17.782	19.443	22.821
13	6.782	8.464	9.470	18.958	20.669	24.139
14	7.477	9.246	10.300	20.128	21.886	25.446
15	8.181	10.036	11.136	21.293	23.097	26.743
16	8.895	10.832	11.976	22.452	24.301	28.030
17	9.616	11.634	12.822	23.606	25.499	29.310
18	10.346	12.442	13.672	24.756	26.692	30.581
19	11.082	13.255	14.525	25.902	27.879	31.845
20	11.825	14.072	15.383	27.045	29.062	33.103
21	12.574	14.894	16.243	28.184	30.241	34.355
22	13.329	15.719	17.108	29.320	31.415	35.601
23	14.089	16.549	17.974	30.453	32.585	36.841
24	14.853	17.382	18.845	31.584	33.752	38.077
25	15.623	18.219	19.717	32.711	34.916	39.308
26	16.397	19.058	20.592	33.836	36.076	40.534
27	17.175	19.900	21.469	34.959	37.234	41.757
28	17.957	20.746	22.348	36.080	38.389	42.975
29	18.742	21.594	23.229	37.198	39.541	44.189
30	19.532	22.444	24.113	38.315	40.691	45.401
31	20.324	23.297	24.998	39.430	41.838	46.608
32	21.120	24.153	25.885	40.543	42.983	47.813
33	21.919	25.010	26.774	41.654	44.125	49.014
34	22.721	25.870	27.664	42.764	45.265	50.212
35	23.526	26.731	28.556	43.872	46.404	51.408
36	24.333	27.595	29.450	44.978	47.541	52.601
37	25.143	28.460	30.345	46.083	48.676	53.791
38	25.955	29.327	31.241	47.187	49.808	54.979
39	26.770	30.196	32.139	48.289	50.940	56.165
40	27.587	31.066	33.038	49.390	52.070	57.348
41	28.407	31.938	33.938	50.490	53.198	58.528
42	29.228	32.812	34.840	51.588	54.324	59.707
43	30.052	33.687	35.742	52.686	55.449	60.883
44	30.877	34.563	36.646	53.783	56.573	62.058
45	31.704	35.441	37.550	54.878	57.695	63.231
46	32.534	36.320	38.456	55.972	58.816	64.401
47	33.365	37.200	39.363	57.065	59.935	65.571
48	34.198	38.082	40.270	58.158	61.054	66.738
49	35.032	38.965	41.179	59.249	62.171	67.903
50	35.869	39.849	42.089	60.339	63.287	69.067

Figure 4. Confidence limits on the mean of a Poisson distribution, given a single sample value, n, of the random variable.

number of trials to get the results as percentages. For example, suppose an experiment with good statistical representation results in 100% accuracy. From the table at 0 errors, the average number of errors per experiment over many repetitions of the experiment should be somewhere between 0.010 and 4.605, with 98% confidence.

Training Sets and Test Sets

A well established empirical fact is that if a pattern recognition machine is tested on the same data base used in training the machine, the results are always better than if an unknown population is employed for the test. The contributions to this bias can be rather subtle, so the safe test procedure involves procurement of a completely new test data base from talkers not previously used in any part of the engineering development project. (On the other hand, development work directed toward a specific application is best done from recordings of real or simulated operational conditions in order to minimize a different kind of bias.)

Surprisingly, there is very little information on this subject in the mathematics literature. Dialog has, therefore, established a modest analytical project to derive expressions for the statistical bias in cases of interest for pattern recognition. One interesting result for maximum likelihood recognition of Gaussian patterns is that the expected test score for an unknown population is pessimistically low when the training set is of finite size. Figures 5 and 6 show the relation between expected training set and test set scores for one dimensional Gaussian patterns. This particular function is of little or no practical value, since all cases of interest are multi-dimensional. The diagrams illustrate, however, that even in this simple case the bias is a complicated function of the population size and the true error rate.

By the law of large numbers, the bias decreases inversely with sample size for a properly designed method. The system behavior as a function of sample size can, therefore, be estimated roughly by taking measurements at two sizes. At every stage of development and field testing, however, the inescapable conclusion is that small scale pattern recognition tests yield very unreliable estimates of large scale performance.

II. Accomplishments and Planned Development at Dialog

Our company, Dialog Systems, Inc., was formed in 1971 for the purpose of developing and commercializing speech recognition equipment. The concept derived from earlier work engaged in at Listening, Incorporated on marine bioacoustics, acoustic signal processing, and psychoacoustics. The original idea passed through well-known stages of theory, experiment, development, lack of financing, financing, sales and is now

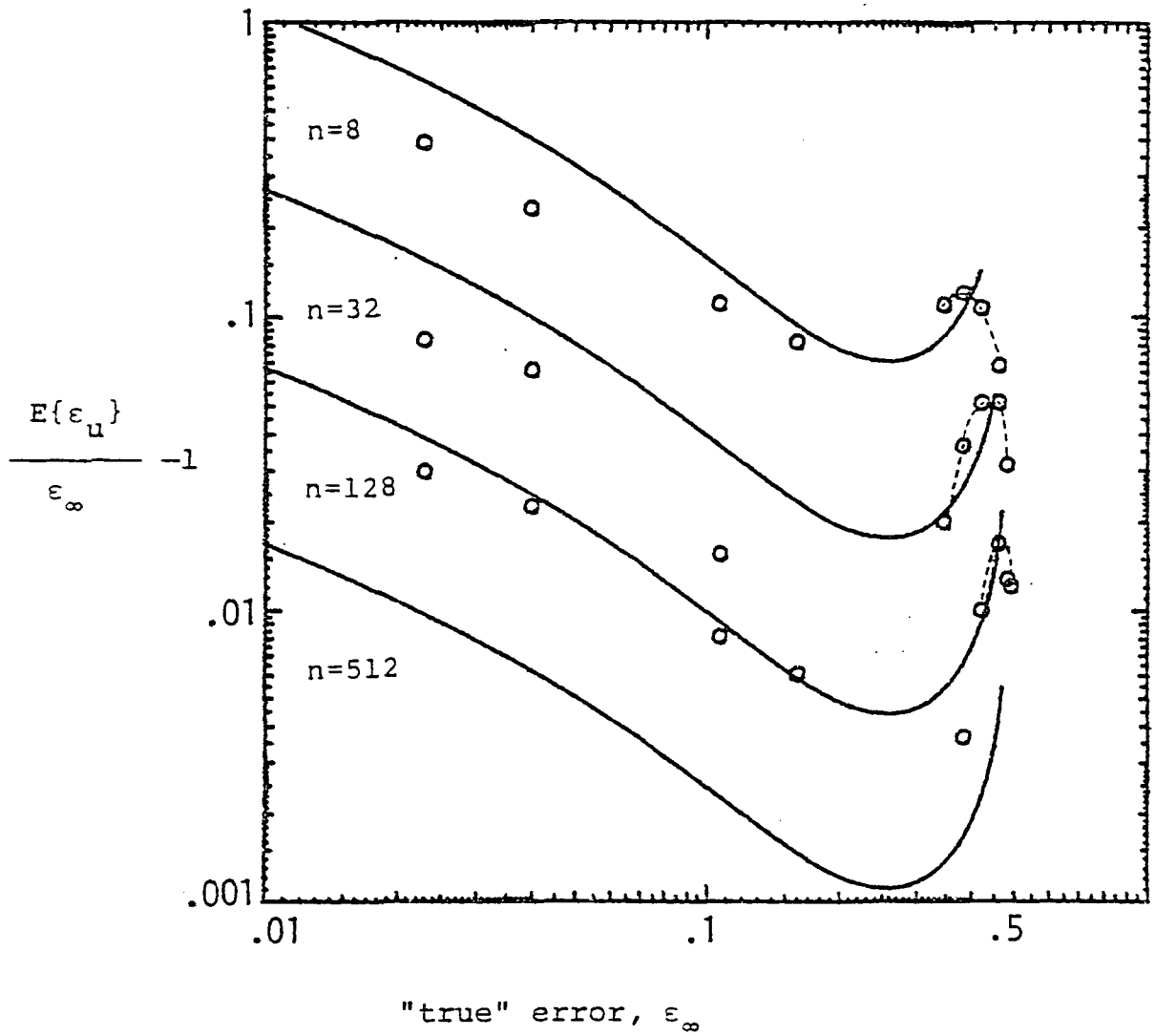


Figure 5. Empirical values for the bias of unknown test set error rates, derived by a computer experiment on pseudo random numbers, compared with a theoretical approximation. The empirical data are relatively accurate (dotted lines) near $\varepsilon_\infty = 0.5$, and the divergence of the Taylor series approximation is evident here. Elsewhere agreement seems rough, but is within experimental error.

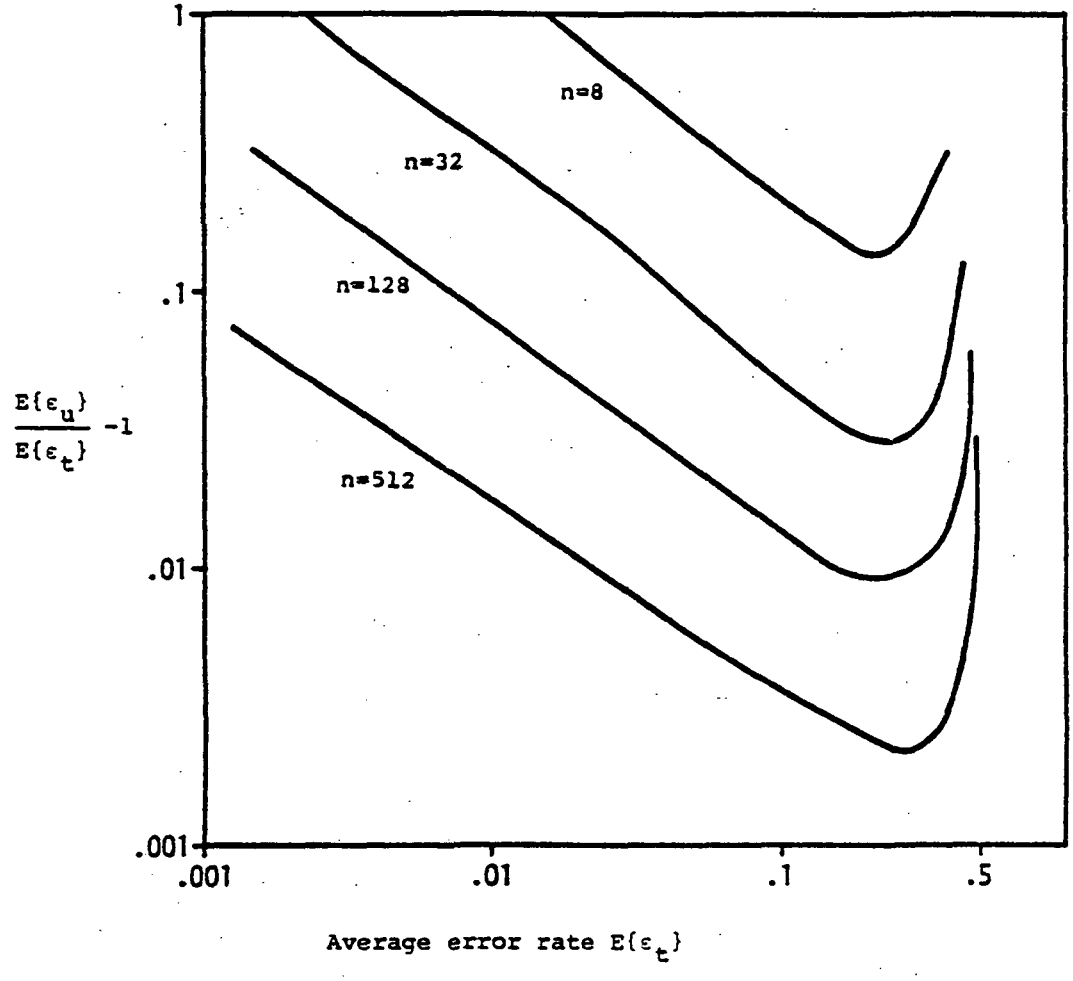


Figure 6. Empirical curves for the relation between average training set and test set error rates for various values, n, of the size of the training set. The functions terminate as indicated near $\varepsilon_t = 0.5$.

The subscript "t" indicates the training set and "u" the unknown test set. The value, ε, is the observed error rate for one experiment of sample size, n.

at the highly advanced state "production engineering headaches". Dialog employs 45, of whom 14 are degreed technical people. The company recently moved from Cambridge to a 20,000 square foot two-building campus complex in Belmont, Massachusetts.

The major product is an eight-channel isolated word system intended for talker-independent switched telephone speech input. With trivial software modification, the same equipment adapts to and tracks each talker's voice characteristics, thus becoming a partially or fully trained machine which is unusually forgiving with respect to changes in the talker's manner of speaking. Operation in the talker-dependent mode requires only one training sample of each vocabulary word. This is made possible by virtue of the precomputed statistical reference patterns contained in the machine.

A complete system (Figure 7) comprises:

1. An analog section consisting of a telephone line switch matrix concentrator, analog-to-digital domain conversion unit and a voice response unit.

2. A disk storage unit for logging and program loading.

3. An interface control computer.

4. A fast signal processing computer of Dialog design and interface to controlled equipment.

5. Power supplies.

The most complex of the units sold to date were priced at about $75,000 for eight simultaneous channels. Installed systems are being supported very heavily by us to ensure that we hear about and correct any troubles encountered. Phone calls from end users are tape recorded and the unintelligible ones analyzed to develop improvements in the recognition algorithm or in the human factors. We have found this operational not-test-but-real-life condition to be different from any simulation, and in the case of new applications to require a substantial refinement effort after delivery and installation. In our experience, problems have arisen that could not be solved by either recognition software or control software changes alone. In general, therefore, the manufacturer must plan for this extra effort, or else the customer must have a speech recognition expert on hand to make his system work. The author has never heard of a speech system working well in an application for which it was not designed, and believes that this situation will continue to be true for some years to come, until a really broad range of application problems has

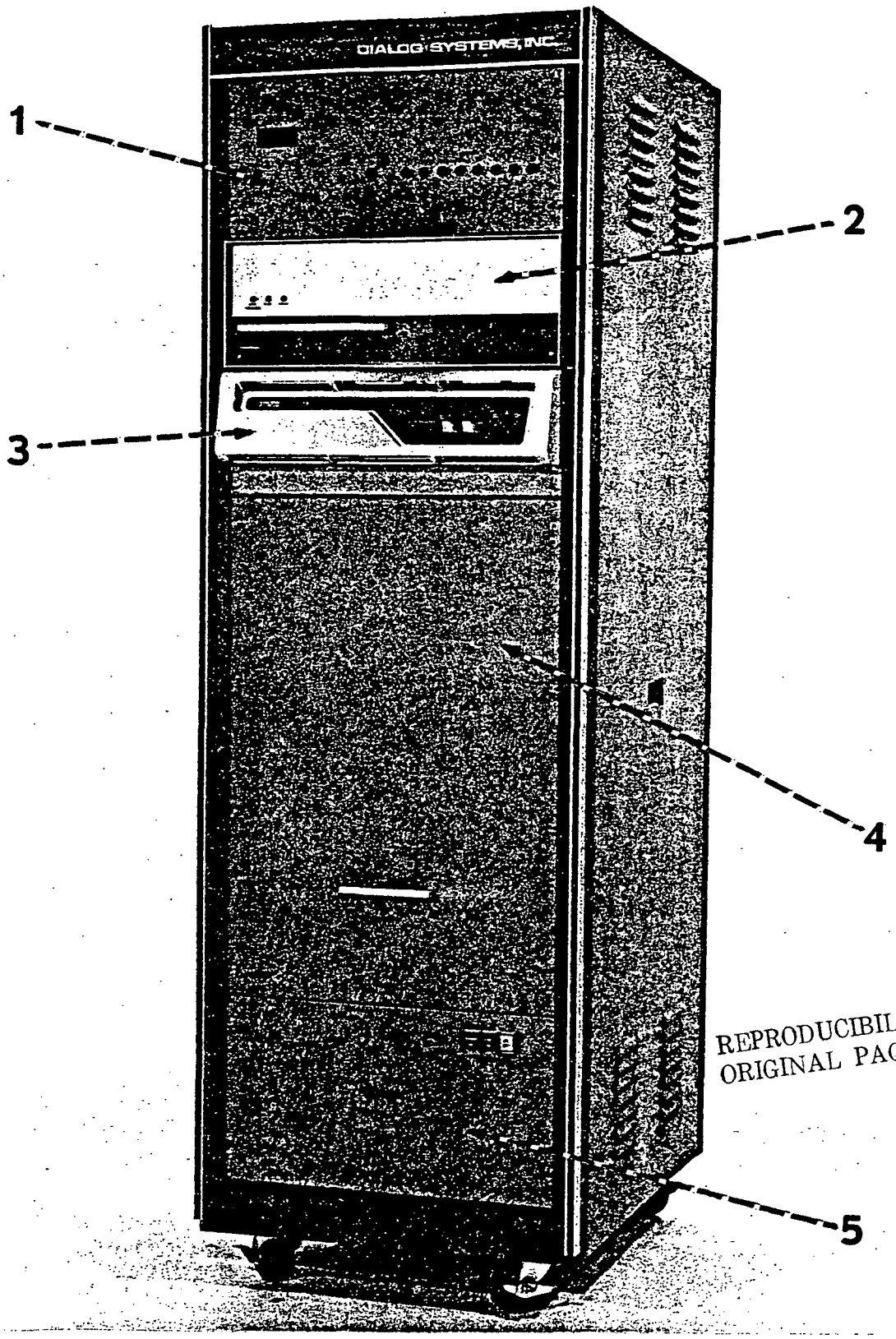


Figure 7. Complete Speech Recognition System

been solved. The problems are being solved, and no one need hesitate to take the next step; but the step is there, cannot be skipped, and it costs money and manpower to take it.

In addition to its heavy investment in product commercialization, Dialog has the resources to maintain a strong research effort; and the company is actively searching for talended people who will follow their own interests in the general area of continuous speech recognition. About 80% of the company's R&D effort is in-house funded, and research personnel (except for the author) are relatively well sheltered from the vagaries of business problems. The main distinguishing feature of the development work at Dialog is that we are making a serious attempt to find improved statistical models of speech data. This includes taking into account the measured variances and cross-correlations of various parameters over large populations of talkers. Thus, we speak of our pattern matching functions as "conditional probability densities" and not "distances". There is, in fact, a fundamental mathematical difference, because a distance is a symmetric relation. Probability measures do not have this property, and do not want it.

Aside from its intrinsically more precise description, an advantage of this approach is that a small number of reference templates suffice for a talker-independent representation. This greatly reduces the workload on higher-level calculations, particularly in talker-independent continuous speech recognition. Our current talker-independent telephone speech product incorporates just two reference patterns per vocabulary word, derived from the speech of hundreds of talkers. The task of gathering, labeling, and proofreading the raw speech data bases for this work has turned into a major project in itself.

While Dialog's engineering activity has so far been devoted to development of system hardware and isolated word recognition, our research effort since 1974 has concentrated on continuous speech recognition. Under contracts with RADC, we have worked on the keyword spotting problem and have produced an algorithm with good talker-independent performance (Figure 8). Word spotting tests under simulated operational conditions are scheduled for 1978. The keyword task is quite difficult, because the brief target sound must be detected independent of context, and all other sounds of an open, plain language input stream must be rejected. The problem is made interesting, however, by the fact that the total number of variables is manageable, so that it is possible to develop theoretical hypotheses and test them by experiment.

The task of limited vocabulary continuous speech has fewer uncontrolled variables than the keyword task, and is therefore easier. Dialog demonstrated such a system in 1975; this system is now in the product development phase and will be released for operational use in 1978.

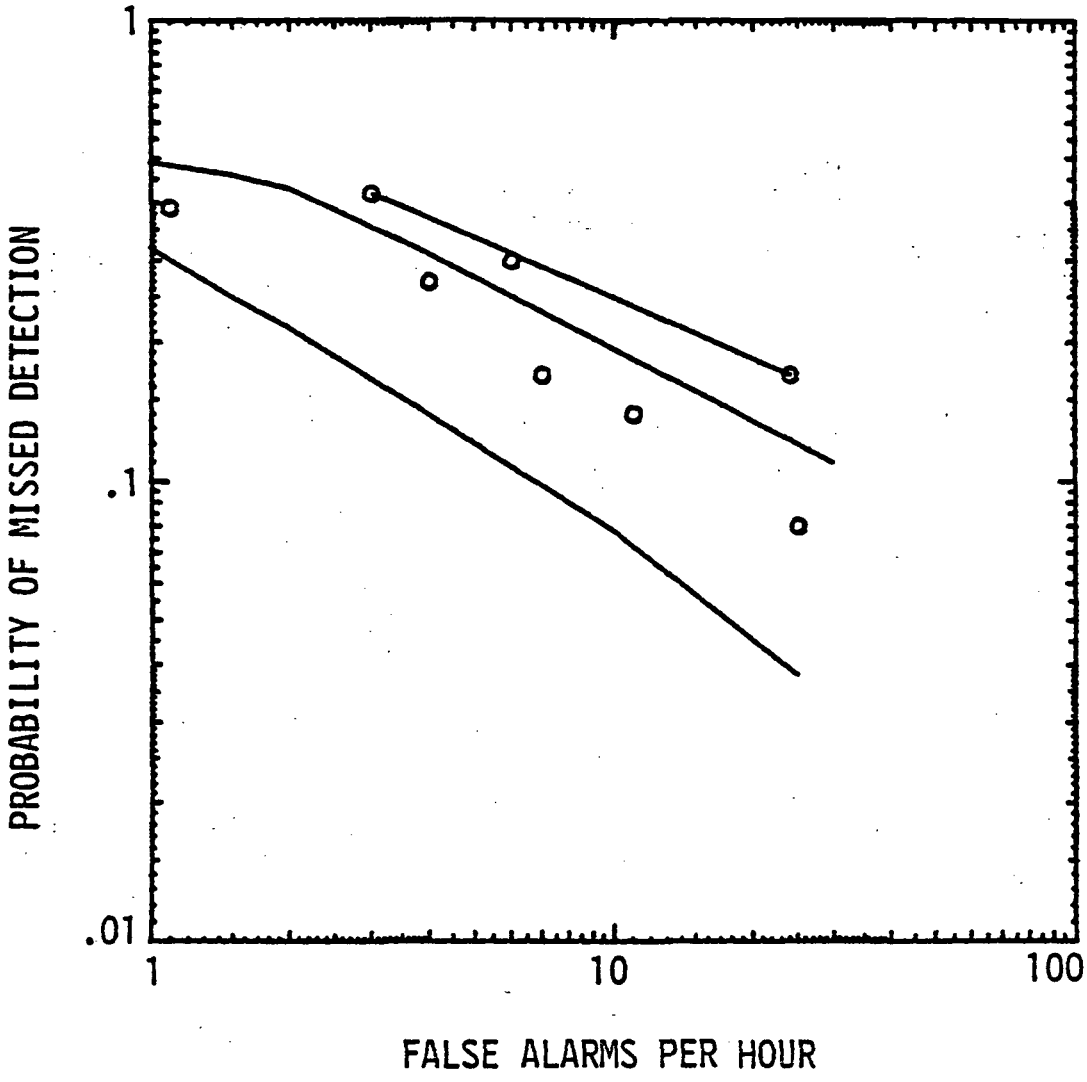


Figure 8. Receiver operating characteristics derived from English language tests of the key word recognition system. Top curve recalls data from the 1976 test, while the two lower curves form an approximation 90% confidence band for the results of the 1977 test.

Organizations which now have operational speech recognition equipment for sale tend to be strong on acoustical pattern analysis and weak on higher level linguistic analysis. There has probably been a general feeling among these people that linguistic processing will not cure the acoustic-level problems on which they feel they are making progress anyway. However, at least two groups, the ones at Dialog and Texas Instruments, have made use of check digits to do error correction on digit string inputs. A series of digits with a check sum is a language; the rule for checking the check digit is a linguistic rule for deciding if a sentence belongs to the language. Thus we already have in practical equipment a rudimentary sort of linguistic processing - and it is not to be scoffed at, because it does reduce the error rate (see Figure 9).

Syntax branching rules have, of course been in use for a long time; but there is still a gap between the well-understood techniques in acoustic analysis and statistical pattern recognition and the realm of linguistic analysis. This gap is being filled in by relatively slow, careful experimental work. It may be expected that practical commercial continuous speech systems with vocabularies of several hundred words will appear in the early 1980's, but probably not within the next two years.

BIOGRAPHICAL SKETCH

Stephen L. Moshier

Stephen L. Moshier is President of Dialog Systems, Incorporated, Belmont, Massachusetts, and is specifically responsible for the direction of its research effort in addition to his general administrative duties.

Mr. Moshier has been with Dialog Systems since 1971 and has made major contributions to that company's practical implementation of computerized speech recognition.

From 1965 to 1971, he served variously as Engineering Vice President, Technical Director and President of Listening, Incorporated, Arlington, Massachusetts, where he worked on the development of special purpose transducers and instruments for speech analysis, animal training, underwater acoustics and spectrum analysis.

Mr. Moshier attended Harvard College (Physics), received a Bachelor of Science Degree (Methematics), Summa Cum Laude, from Boston University in 1971 and did graduate work in communications biophysics at M.I.T. in 1971-1973. He has published many papers and patents in the field of speech recognition.

Mr. Moshier is married, has one child and resides in Cambridge, Massachusetts.

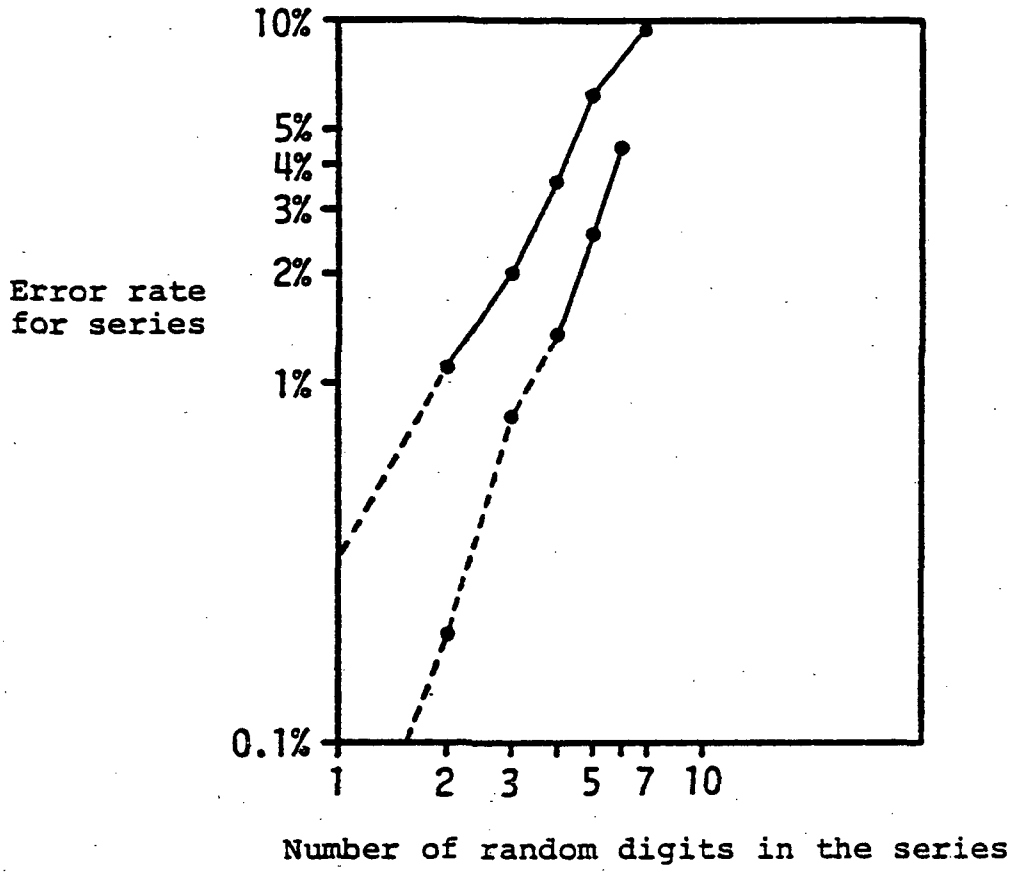


Figure 9. Observed average recognition error rates for random series of isolated spoken digits having check digit(s) appended. For each input series the computer chose the series possessing the best cumulative likelihood score and also having legal check digits. The curves show speaker-independent performance for 25 male voices recorded over standard switched public telephone connections. Error rates below 1% were statistically unreliable in this experiment. No errors were observed for series of length 1.

Upper curve: one check digit adjoined to the indicated number of random digits. The check digit is the 9's complement of the modulus 10 sum of the random digits.

Lower curve: two check digits adjoined. The first check digit is the same as above. The second check digit is the 9's complement of the modulus 10 sum of the squares of the random digits.

DISCUSSION

MR. Robert Osborn

Q: Arnold Popky, Threshold Technology: You talked about voice recognition over the phone lines. I was wondering if you felt any of the constraints of the bandwidth of the phone lines and the microphones supplied by the telephone company on your recognition accuracy.

A: That is an area that we have good confidence in at this point. Our algorithm is made purposely transparent to that a limited bandwidth which was never a problem, we never did rely on wide bandwidth, high fidelity speech. And the dynamics of the microphone are adequately taken care of by the statistical approach used. The actual data base is collected. It's a tremendous job to collect adequate data base. We have a data base department that does nothing but record voices, label them, and digitize them. It consists of five full time people. They just collect voices. And it's not easy to collect voices over the random telephone lines. It takes a good deal of effort. We now have data bases consisting of many hundreds of people -- individuals speaking over random telephone lines. It took quite a while to get that. The base recognition accuracy that you saw on the slide for the dialogue speaker independent system, was telephone speech. It was over telephone lines.

Q: Michael Nye: You explained two applications, one was a radio paging application, and the other application for telephone switching. I have two questions. One is, it wasn't clear to me, the benefit that speech offered in that application. And I was wondering if you could just comment for 15 seconds about what that is, why is speech used in that application. And secondly, you outlined a standard system configuration and showed a picture of a device. What would an 8 terminal system like that typically cost, if you can quote a number like that.

A: The answer to the second question is somewhere in the range of $80K. The first question, well, what other types of solutions are there? Yes , you can have a bank of operators listening to people, you can have people touchtone things, or you can have N telephone numbers for N number of people with pocket pagers. Each of those has several economic or operational problems. In the case of touchtones, they don't exist extensively. Installed touchtone base in New York is maybe 15 or 20 percent for instance. In some places they don't have touchtone at all, effectively, Touchtone pads don't seem to work adequately, and they're an additional capital expenditure, much more so than the system that we've presented. Operators are very expensive, especially in places where labor costs go out of sight. Our commercial

is in Canada and they're locked-in because they can't get telephone numbers from the telephone company, and if they could, the telephone company would charge them $30 a month for them. And that five or ten thousand subscribers adds up. This is a true economic application area; somebody wrote down on the balance sheet what the results would be, and they came out with voice, I think that's the way we have to approach a lot of these application areas, we've just got to solve the problem.

Q: George Doddington, TI: First, the simple question, what were the key words that were used in that plot of key word performance?

A: That's available in the Rome report. There were a number of key words tried, not only in English.

Q: George Doddington, TI: Second is a comment, that is that, you mentioned a little formula of x^n for performance as a function of vocabulary, and I don't really disagree with that. But I think I would like to make the comment that the performance depends more on the vocabulary than it does on the vocabulary size, and as an example, I would say that we at TI have done some work on nested vocabularies, say, from 100 to 800 words, and we've found, for example, that the performance on the 100 word vocabulary, which are most commonly used words, is poorer than the performance on the 800 word vocabulary, which includes the 100 word vocabulary.

A. We saw that with the AMES group, too. Also another fact that seems to be rediscovered constantly is that the errors are very heavily concentrated on some speakers, they are not uniformly distributed over all speakers, I don't think there is an adequate explanation of why tha's true. We're certainly investigating it. It's an observation I believe other people have made at times, too. It's not related to stress, necessarily.

Q: Ed Huff, NASA Ames: How do you account for the fact,I guess, that your 800 words is dealt with more competently that the 100 word subset.

A: George Doddington, TI: The 800 words are dealt with in exactly the same way as the 100 words, it's just that the extra 700 words that you throw in are more easily recognized. There's this shorthand principle I guess, in speech, that the more often you use a word the shorter it tends to get over eons of time of language evolution, so that the most commonly used words in English are one syllable words like "the," "of," "and," so for example, take the first 100 words. It may be an average of one and a half syllables per word. But after you get beyond the first several hundred words the average number of syllables per word is up around two, and, as everyone knows, two, three and four syllable words are very easy to recognize; the problem is with one syllable words.

Q: Ed Huff: So in other words, you're taking advantage of apertures in order to obtain that result.

A: George Doddington, TI: Oh, yes. We count on the fact that in the exercise of the 800 word vocabulary the first 100 words account for only one-eighth of the exercise. Even if you weight the first 100 words according to frequency of usage, given that they're used more, the results are still the same.

Q: Rex Dixon, IBM: I think one of the things that's happening here is really unfortunate, and that is that we're tending to go to generalizations. For example, the generalization of difficulty of recognition as vocabulary gets larger, with no conditionals, which, of course, as you've pointed out, George, this is a misleading statement at best. I think also your statement about as words get longer they get easier to recognize, is also a generalization. I think any of us, who have been in the speech area, can come up with a vocabulary of long words that will be extremely difficult to recognize, that is to get accuracy within that list, and at the same time come up with a set of very short words that are very easy to recognize. So I think the thing we need to do here is to stop this perpetration, or perpetuation of over generalizations which keep the field in trouble all the time. People go around saying, "Well, as the vocabulary gets bigger, it gets harder to recognize"; "longer words are easier to recognize than short words," etc, etc. And it just simply isn't true. I mean, these things are all conditioned by a lot of other variables. Now the thing we should be about, relative to vocabulary and difficulty of recognition, is saying things like, "here is a method by which you can calculate, using what we know about difficulty, having to do with phonetic similarity, with vocabulary size, here is a way of predicting the difficulty of a particular vocabulary, using all these factors." This is the basic research I hope you were referring to. If we had these things, I think the task for application selection would made easier.

A: I think that's correct.

Q: Jared Wolf, BBN: Just to go along with Rex's statement, I'd just like to point out for some people that may not be familiar with it, that there was a thesis in Carnegie Mellon by Gary Goodman last year which by no means is the whole attempt, but it's a very good first start in just the direction that Rex just mentioned. People should be well aware that we're looking for applications. I don't think it takes care of everything, but it's a lot better than nothing.