research memorandum RJ 168

IBM RESEARCH LABORATORY SAN JOSE, CALIFORNIA May 2 7, 1959

SUPPLEMENTARY NOTES ON OPTIMUM BLOCK LENGTH FOR DATA TRANSMISSION

PART I

F. B. WOOD

SUPPLEMENTARY NOTES ON OPTIMUM BLOCK LENGTH FOR DATA TRANSMISSION

PART I

by

F. B. WOOD

ABSTRACT

This is a supplement to Report RJ-MR-11 "Optimum Block Length for Data Transmission with Error Checking." Data on the error probability in an experimental data set with impulse noise interference are put into the theoretical formula from Report RJ-MR-11 for optimum block length to determine a lower bound on the optimum block length. These results constitute a lower bound, since the noise tape used in the experiments was recorded on a noisy telephone line. The lower bounds on optimum block length for this experimental IBM Data Set are 2700, 1800, and 1000 characters respectively for transmission rates of 1000, 1600, and 2400 bits/sec at a 0 db signal-to-noise level. The IBM Data Set (phase modulation) is compared with the AT&T Co. Subset (frequency modulation) at a signal-to-noise level of -5 db and a transmission rate of 600 bits per second.

TABLE OF CONTENTS

SUPPLEMENTARY NOTES ON OPTIMUM BLOCK LENGTH FOR DATA TRANSMISSION

Introduction

Previous papers on optimum block length dealt only with theoretical cases or used arbitrary assumptions as to the error probabilities in computing examples of optimum block length.[1,2] In the revised version of reference 1, a set of curves was substituted in place of the assumed values so that one could insert experimental data on the error probabilities at a later date. By the time of the oral presentation of reference 2, E. Hopner had presented, at the Eastern Joint Computer Conference, some experimental curves of bits per error versus signal-to-noise ratio for an experimental Data Set.[3] Therefore, Figure 8 of reference 2 was deleted in the oral presentation, and sample calculations using E. Hopner's data on the IBM Data Set (phase modulation) were substituted. The Bell System has an experimental Data Subset (frequency modulation).[4] Comparable error data have been taken on both the IBM and the Bell System Data Sets. In view of a policy of not releasing data on the error rates of other companies' equipment to the engineering public, only the error rates on the IBM Data Set were released at the AIEE discussion. Therefore, this version of the paper, which includes error rates of the A. T. &T. Co. Subset, is intended for internal company distribution only.

Preliminary data are available on tests made on other telephone lines by A. T. &T. Co. with the A. T. &T. Co. Data Subset.[9] Some of the other lines tested by A. T. &T. Co. have noise distributions with a much higher percentage of multiple errors than the noisy line used for our tests. These data are not used at this time, because A. T. &T. Co. expects to have a more extensive analysis of error statistics available in a few months.

This report is a supplement to Report RJ-MR-11, "Optimum Block Length for Data Transmission with Error Checking," and is designated as "Part I" to distinguish it from future supplements which may be published when the A. T. & T. Co. noise statistics are available.

The function of this report is twofold: (1) to document the additional optimum block length data released in the oral discussion at the February 2, 1959 AIEE meeting, and (2) to make available equivalent calculations for the Bell System Subset for distribution within IBM. The data on error rates for both IBM and Bell System Data Sets were made with the same noise tape.

Optimum Block Length

The definition of optimum block length used in this study is the number of characters per message block which would maximize the transmission efficiency. Since this study does not include the analysis of the economic costs of the channel, buffer storage, and logic, it does not necessarily follow that one should operate at the optimum block length.

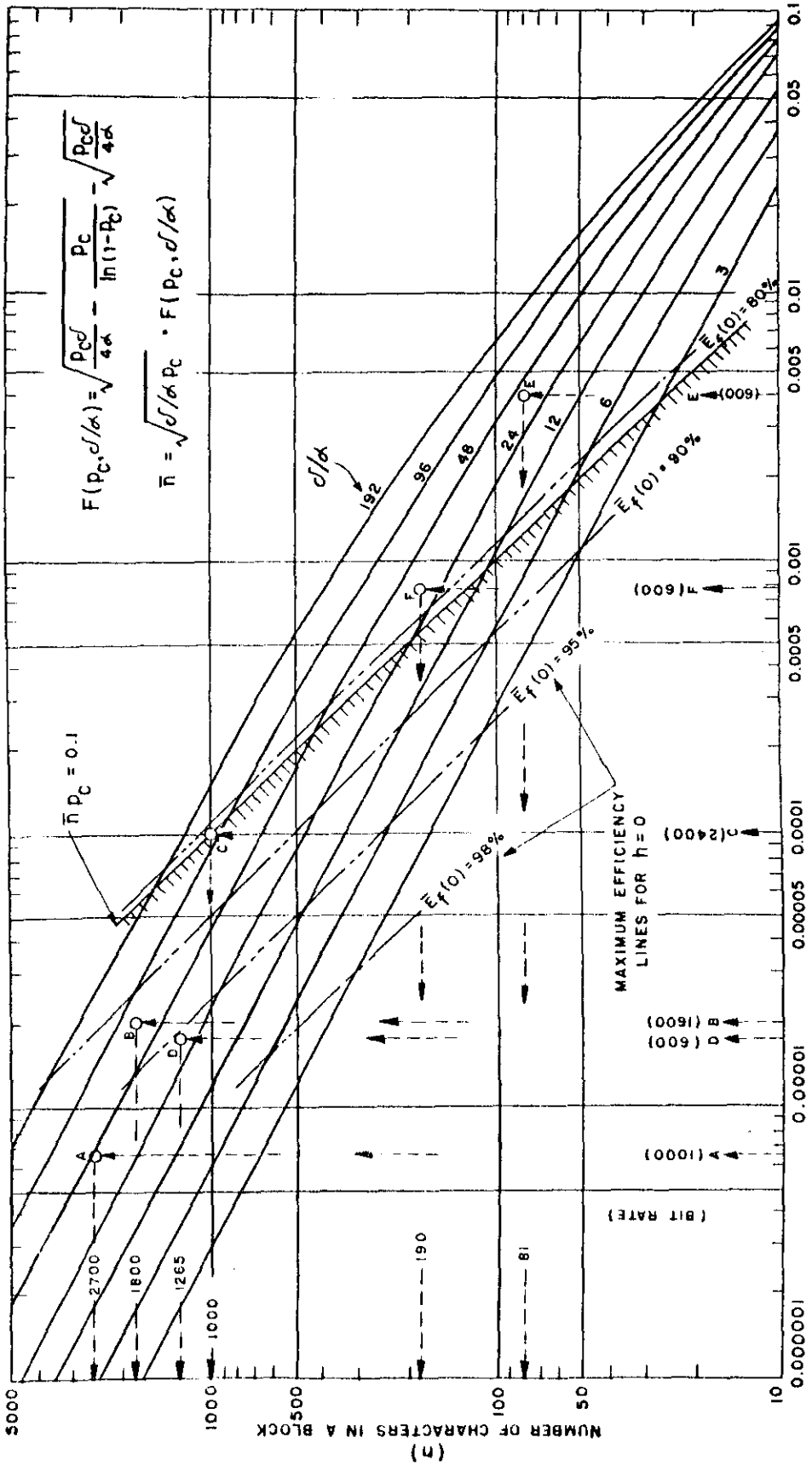

Figure 1. Optimum Block Length vs. Probability that a Character is in Error.

For those interested in pursuing the more complete problem, the work of A. Hauptschein, L. S. Schwartz and others at New York University* is of significance.[5]

Since the experimental data reported by E. Hopner extend to lower error probabilities than were estimated when the curves for references 1 and 2 were prepared, the curves from reference 2 (Fig. 2) are extended in Fig. 1 of this report. The sample calculations made in this report are shown graphically in Fig. 1.

Mr. J. B. Norris of Poughkeepsie has conducted independent studies of optimum block length using a different approach to the problem. **

Experimental Error Probability

Some error data is available for teleprinter lines from reports of the Royal Aircraft Establishment, Farnborough, England.[6, 7] The transmission speed involved in these teleprinter links was too slow to make optimizing the block length of real importance. General statistics on error rates with another modulation-demodulation system (FM Data Subset) are being accumulated by American Telephone and Telegraph Co. and Bell Telephone Laboratories. Until A. T. &T. Co. has sufficient data analyzed for release to industry, we are limited to using a single tape recording of a twenty-five minute sample of a "noisy" line. Even when A. T. &T. Co. data is available it will be limited by the modulation system used in the experiments. E. Hopner has reported on tests of an experimental high speed data set in which this tape recording was used to introduce impulse noise. The curves of Fig. 2 show the variation in error rate as the relative signal level was changed to simulate different signal-to-noise ratios. The random (thermal) noise curves of Fig. 2 are not used in this analysis, because one would normally operate at a much higher signal-to-noise ratio than the range for which the random (thermal) noise errors are significant.

The dotted lines in Fig. 2 are for tests on the preliminary Bell System Subset conducted by IBM in San Jose with the same noise sources as were used in testing the IBM Data Set. One difficulty in comparing the tests on the two Data Sets is that the 600 bit/sec tests on the IBM Data Set could not be extrapolated to the zero db noise-to-signal level, so it was necessary to compare the two sets at the 5 db level, points D and E in Fig. 2. Since the 5 db level represents a higher noise level than would be encountered in practice, an additional point F was calculated for the Bell System Subset at the zero db level in an attempt to come closer to the true operating conditions. Furthermore, it should be remembered that the Bell System Subset is not their final design, but a preliminary model.

* The reports I have examined so far from this group relate to a region of block length an order of magnitude smaller and also compare the "decision feedback" considered here with other possible systems.

** The definitions and resulting curves prepared by Mr. Norris have been compared with the definition and curves of reference 2 and have been found to be in agreement when one system of definitions is translated to the other.

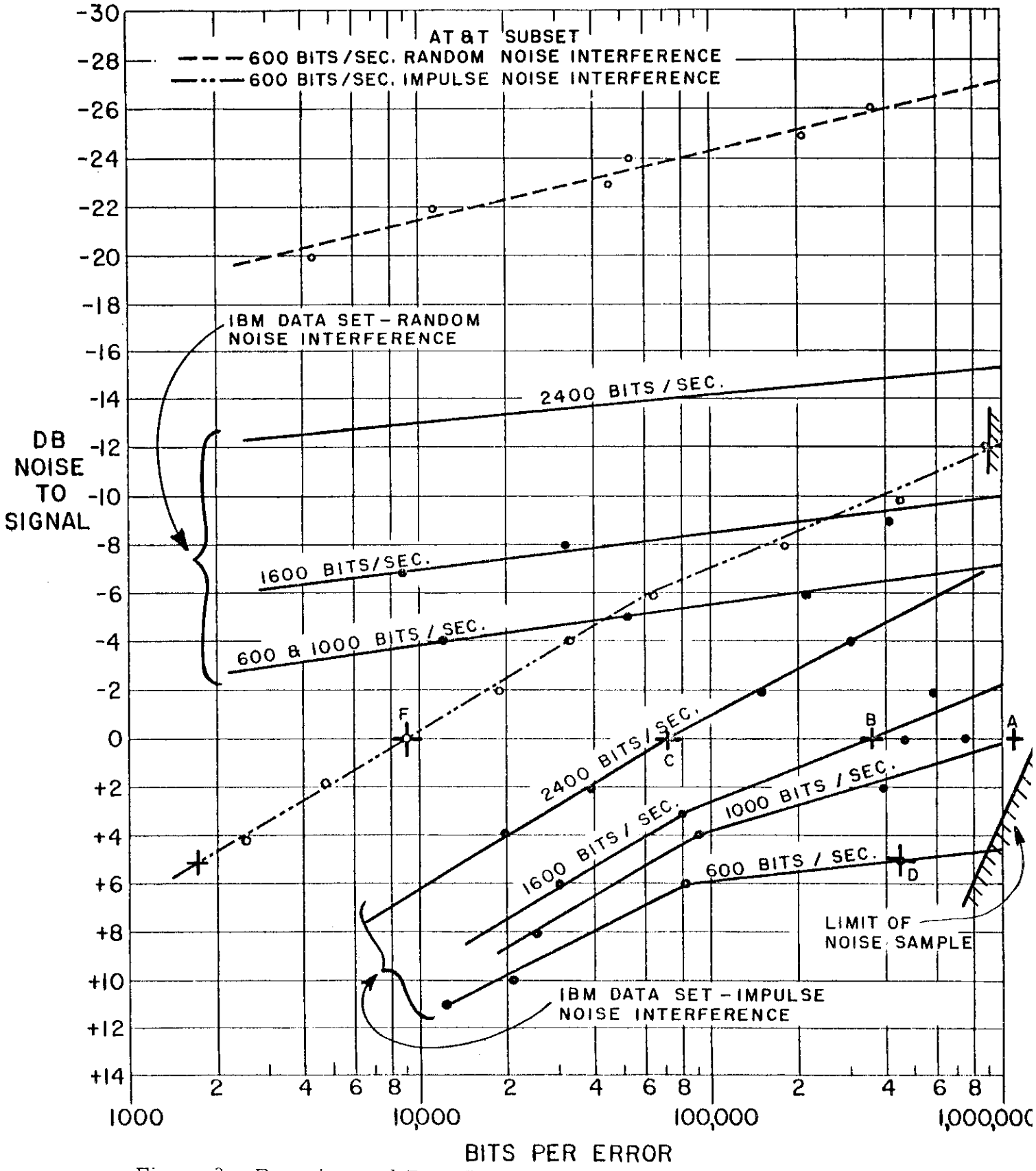

Figure 2. Experimental Data Set Error Rates.

The intersections of the curves for 1000, 1600, and 2400 bits/sec with the zero db noise-to-signal line, marked A, B, and C in Fig. 2, are the sample values of error rates used in this analysis for obtaining optimum block length as a function of transmission rate for the IBM Data Set.

Sample Determination of Optimum Block Length

The values of B (bits per error) for zero db noise/signal ratio* for the IBM Data Set for 1000, 1600, and 2400 bits/sec are respectively 1.1×10^6, 3.5×10^5, and 7×10^4. The values which are obtained graphically from Fig. 2 are tabulated for future reference in Table I. Averaged values from the smoothed curves in Fig. 2 were used to obtain these. They are marked as points A, B, and C in Fig. 2. These error rates are for a noisy 100-mile-long N-1 channel. By assuming the errors are independent, we have the probability that a bit is in error:

$$p_b = (1/B) \tag{1}$$

This gives p_b equals 9.1×10^{-7}, 2.86×10^{-6}, and 1.43×10^{-5} respectively. Assuming a seven-bit code including one redundant bit from Appendix III of reference 2, where "a" is the number of bits per character, and assuming independent errors, the probability that a character is in error is:

$$p_c = \sum_{i=1}^{a} \binom{a}{i} p_b^i (1-p_b)^{a-i} = 1 - (1-p_b)^a$$

$$= 1 - \left(1 - a p_b + \frac{a(a-1)}{2!} p_b^2 - \ldots\right) \approx a p_b, \tag{2}$$

for $p_b \ll (a-1)/2$.

Review of Appendix III of reference 2 shows that for $p_b < 0.00014$, the first terms of equation 2 is a reasonable approximation:

$$p_c \simeq 7 p_b \tag{3}$$

This gives in the same order for the respective cases: $p_c = 6.36 \times 10^{-6}$, 2.0×10^{-5}, and 1.0×10^{-4}. These three values have been marked on Fig. 1 and are labelled with the corresponding bit rates.

---

* This "noise/signal ratio" is the inverse of the "signal-to-noise ratio" for gaussian thermal noise (random noise). For impulse noise this relationship is not valid, because the detailed distribution of the impulse noise is not known, and the available impulse noise probability curves are given in terms of a counting of peaks which cannot be rigidly defined without knowing the spectrum.

The next step is the graphical determination of $\bar{n}$ and E as is done in the second section of Appendix III (reference 2).

Step (1): The time per character, $\alpha = 7/f$. (4). For the three cases α is respectively: 0.007, 0.0044, and 0.0029 (sec/char).

Step (2): Consider the case of a transcontinental voice circuit where the limiting factor in the delay time is the 0.300 (sec) allowance for setting and resetting the echo suppressors. Some equipment has been designed for twice this delay or 0.600 sec to allow for propagation time and setting and resetting of echo suppressors. The longer time may be necessary to allow for all possible dial-up circuits in the United States. In this report the 0.300 sec figure has been retained to keep the results consistent with sample calculations in reference 2. Assume that two characters are used for block-check and answer-back signals and that the logical operations can be completed within one character time* so that the number of character times of the delay is:

$$\delta/\alpha = 3 + (0.300/\alpha). \qquad (5)$$

For the three cases, then δ/α is respectively 47, 71 and 106. These points are interpolated and marked as points A, B, and C in Fig. 1.

Step (3): On Fig. 1 the intersection of curve $\delta/\alpha = 47$ with $p_c = 6.36 \times 10^{-6}$ is marked "A." This gives:

optimum block length, $\bar{n} = 2700$
maximum efficiency, $E = (6/7)\ 96 = 82\%$

The results from above and the other points taken graphically from Fig. 1 are tabulated in Table I.

Examination of Table I shows that for independent errors the lower bound on the optimum block length for the IBM Set ranges from 2700 characters for 1000 bits/sec to 1000 characters for 2400 bits per second.

Similar calculations were made at 600 bits/sec, 5 db level for both the IBM and the Bell System Data Sets. These results are drawn in graphically in Fig. 1 and are also tabulated in Table I.

* For example, in N. M. Abramson's report on SEC-DAEC codes, the checking number which indicates whether an error occurred appears in the checking register when the last redundant bit of the character is processed.[8]

TABLE I

OPTIMUM BLOCK LENGTH

& EFFICIENCY

Experimental Data Sets on Noisy Lines With 7-Bit Code

Data Set	Transmission Rate (bps)	S/N (db)	B Bits per Error	Delay $\frac{\delta}{\alpha}$	p_c	Lower Bound on Opt. Block Length $n_{opt.}$	Efficiency of Transmission	Point	Net Information Rate (bps)
IBM	1000	0	1.1×10^6	46	6.36×10^{-6}	2700	96x6/7 = 82%	A	820
IBM	1600	0	3.5×10^5	71	2.0×10^{-5}	1800	96x6/7 = 82%	B	1310
IBM	2400	0	7×10^4	106	1.0×10^{-4}	1000	82x6/7 = 70%	C	1680
IBM	600	-5	4.4×10^5	29	1.8×10^{-5}	1265	82x6/7 = 70%	D	(420)*
AT&T	600	-5	1.8×10^3	29	3.9×10^{-3}	81	54x6/7 = 46%	E	(258)*
AT&T	600	0	9×10^3	29	7.8×10^{-4}	190	74x6/7 = 64%	F	384

* Note that these two points are for a 5 db higher noise level than the other points.

Efficiency and Net Information Rate

At 1000 bits/sec the IBM Data Set has an efficiency of 82 percent so the net information rate is 820 bits/sec. As the transmission rate is pushed up to 2400 bits/sec the efficiency drops to 70 percent with a net information rate of 1680 bits/sec. These values are also tabulated in Table I. At 600 bits/sec and 5 db noise level, the IBM Set has a net information rate of 420 bits/sec. One would not normally operate at this high noise level. The use of data at this noise level is for the purpose of comparing the experimental IBM Data Set and the preliminary Bell System Subset at the same noise level. It is not possible to extrapolate the data of Fig. 2 beyond the lines marked "Limit of Noise Sample." To obtain data on the IBM Data Set at 600 bits/sec for higher noise levels would require a noise sample larger than the 25-minute tape.

Significance of the Lower Bound

Let us examine our assumptions to see why these values of optimum block length are lower bounds. The two important factors are: (1) the assumption of independence, and (2) the choice of a noisy line for the experimental tests.

For part of the range of error probabilities under consideration the approximate formula for optimum block length from reference 2 (eq. 16) is applicable:

$$\bar{n}' = \sqrt{8/\pi}\, p_c \qquad (6)$$

In the appendix it is shown that the assumption of independence has a negligible effect in this case. In general, assuming independence when the errors are dependent results in a larger calculated probability of a character being in error that would actually result from the noise distribution. This large calculated probability of error, when inserted in equation 6, gives a smaller $\bar{n}'$ than the real optimum. Since both assumptions give calculated values of optimum block length which are smaller than the true optimum, the assumptions and formulas of this report give a lower bound on the optimum block length.

The noisy line used in these calculations could possibly have a noise probability curve of ten times the typical case. This would mean that the p_c used in these sample calculations may well be ten times too high, making the optimum block length calculated here too low by a factor of three.

Conclusions

The lower bound on optimum block length for the experimental phase modulation system reported by E. Hopner is respectively 2700, 1800, and 1000

characters of seven bits each at transmission rates of 1000, 1600, and 2400 bits/sec for a long line where the principal delay time is the time for setting and resetting the echo suppressors.

Comparing the IBM and Bell System Data Sets at 600 bits/sec at the 5 db noise level gives optimum block lengths of 1265 and 81 characters respectively. If the Bell System (A. T. & T. Co.) Subset is operated at zero db noise level the lower bound on the optimum block length would rise to 190 characters.

The assumption of independence in translating the number of bits per error to the probability that a seven-bit character is in error gives a negligible correction to these sample calculations. The use of a "noisy" line as the standard in counting experimental errors may give a probability of a bit being in error that is of the order of magnitude of ten times the average value.

Therefore, the lower bound in optimum block length calculated in this analysis is the order of magnitude of one third of the optimum block length for an average line. Since we do not have statistical data on the noise distribution on telephone lines except for the one noisy line, the above estimated correction to the "lower bound" is only an educated guess at this stage.

Appendix I

Experimental Distribution of Multiple Errors

In this analysis the probability of the occurrence of multiple errors within a seven-bit character time was assumed to be independent. The original data from which Fig. 2 was plotted has been examined carefully to determine the validity of this assumption. The distribution of multiple errors for both the IBM and the Bell System Data Sets is tabulated in Table II. The zero db noise/signal level was used in this report except where the 5 db level was needed to get comparable points. Examination of Table II shows that for all bit rates except 2400 bit/sec, all errors are single errors for the IBM Set at zero db level and at 2400 bits/sec 96.2% of the errors are single errors. Therefore, the maximum error in p_c due to assuming independence is 4%, and its maximum effect on the optimum block length is - 2%.

Examination of Table II shows that for higher noise-signal rates (lower signal-to-noise) for the IBM Set, the multiple errors occur at lower transmission rates and hence it becomes more important to take them into consideration. Different distributions of multiple errors could be expected from different modulation-demodulation systems.

NOMENCLATURE

a	=	number of bits per character
B	=	bits per error
$E_f(h)$	=	transmission efficiency with h redundant bits per character
$\overline{E_f}(h)$	=	maximum transmission efficiency obtainable when optimum block length is used.
f	=	transmission rate in bits per second
h	=	number of redundant bits per character
i	=	summation index indicating number of bits or characters in error
m	=	number of information bits per character
n	=	number of characters per block
$\overline{n}$	=	optimum number of characters per block
p_b	=	probability that a bit is in error
p_c	=	probability that a character is in error
α	=	time interval of one character
δ	=	reply delay time interval
δ/α	=	number of characters which could be transmitted during the reply delay time interval.

REFERENCES

1. "Optimum Block Length for Data Transmission with Error Checking," F. B. Wood. IBM Report RJ-DR-532-016, September 20, 1957. (Revised version issued as IBM Report RJ-MR-11, February 28, 1958.)

2. "Optimum Block Length for Data Transmission with Error Checking," F. B. Wood. Communications and Electronics, No. 40, January 1959, pp. 855-861. (Note preprint No. 58-1181 has a different numbering of the figures.)

3. "An Experimental Modulation-Demodulation Scheme for High-Speed Data Transmission," E. Hopner. IBM Journal of Research and Development, Vol. 3, No. 1, pp. 74-84, January 1959.

4. "A Frequency-Modulation Digital Subset for Data Transmission over Telephone Lines," L. A. Weber. Communications and Electronics, No. 40, January 1959, pp. 867-872.

5. "Optimum Information-Acquisition System," B. Harris, A. Hauptschein, and L. S. Schwartz. Operations Research, Vol. 6, July-August 1958, pp. 516-529. See also other recent papers from this group.

6. "Transmission of Data over A Teleprinter Link," A. Cohen. Royal Aircraft Establishment (Farnborough), Ministry of Supply, London, W.C. 2, Technical Note; T.D. 2, January 1956. (ASTIA No. AD 9 1919; IBM Library No. 5057.)

7. "A Data Transmission System for Punched Cards," E. S. Mallett and H. W. P. Knapp. Royal Aircraft Establishment (Farnborough) Ministry of Supply, London, W.C. 2, Technical Note: T. D. 25, December 1957. (ASTIA No. AD 157342.)

8. "A Class of Systematic Codes for Non-Independent Errors," N. M. Abramson. Stanford Electronics Laboratories, Technical Report No. 51, December 30, 1958.

9. Memorandum by L. A. Tate of IBM Product Development Laboratory, Poughkeepsie, on a talk by Mr. Gryb of A. T. &T. Co. given at RETMA Subcommittee A6.3. Subject: Preliminary Performance Tests on High Speed Data Transmission. (October, 1957.)

Acknowledgements

I wish to express my appreciation to Mr. E. Hopner and Mr. Orman Meyer for obtaining the experimental data used in the sample calculations.