Research Report

REDUNDANCY REMOVAL IN BINARY SOURCES

FINAL REPORT

prepared for

DEPARTMENT OF COMMUNICATIONS

COMMUNICATIONS RESEARCH CENTRE

by

Stafford Tavares
Queen's University
(Principal Investigator)

August, 1976

Queen's University at Kingston
Department of Electrical Engineering

REDUNDANCY REMOVAL IN BINARY SOURCES

FINAL REPORT

prepared for

DEPARTMENT OF COMMUNICATIONS

COMMUNICATIONS RESEARCH CENTRE

by

Stafford Tavares
Queen's University
(Principal Investigator)

August, 1976

REDUNDANCY REMOVAL IN BINARY SOURCES

FINAL REPORT

prepared for

DEPARTMENT OF COMMUNICATIONS

COMMUNICATIONS RESEARCH CENTRE

by

Stafford Tavares

Queen's University

(Principal Investigator)

Research Assistant:

1. Mr. K.-C. Fung

Scientific Authority: Dr. W. Sawchuk

Contract No. 0SU5-0141

AUG., 1976

TABLE OF CONTENTS

INTRODUCTION

In this report, we continue our investigation of the compression of binary information sources using rate - $\frac{1}{2}$ convolutional codes. In the preliminary report, we provided some of the background which encourages one to investigate data compression by this means. Our early simulation results were promising and further simulations have been conducted. Initially, we used first, second and third order binary Markov sources. Although these provided us with some insights, it was decided that we should restrict ourselves to simpler sources so that we can relate our results to current theoretical investigations. The source which met this requirement is the Binary Symmetric Markov Source (BSMS).

The Binary Symmetric Markov Source

The Binary Symmetric Markov Source (BSMS) is a first order binary Markov source with the extra restriction that it be symmetrical. The state transition diagram is shown in Figure 1.

It has two states, namely, "0" and "1", and it remains in a given state with probability p and makes the transition to the other state with probability (1-p). Because of the symmetry, zeros and ones are equally likely in the steady state. The entropy of a BSMS is given by H(p) where H(p) is the entropy function, i.e.,

$$H(p) = -p \log p - (1-p)\log(1-p)$$

where the logs are to the base 2 and the entropy is measured in bits/binary digit. When p = 0.5, the BSMS reduces to the Binary Memoryless Source (BMS) with equiprobable symbols.

It may be of interest to note that for even such a simple source as the BSMS, the rate distortion function is not known. Currently, only upper and lower bounds are available and some of the work on the lower bounds is very recent. We will briefly describe one of the upper bounds.

The most trivial upper bound is obtained by throwing away the memory between successive source symbols to obtain the BMS with equiprobable symbols. Note that for p = 0.5, this always increases the entropy since $H(p) \leq H(0.5) = 1$. The rate distortion function for the BMS is known and can be evaluated at rate R = 1/2. A tighter upper bound on the BSMS is obtained by grouping the output digits in pairs and then ignoring the memory between the pairs. If we let $R_2(D)$ be the rate distortion function for this source and $R(D)$ for the BSMS, then it can be shown that [1]

$$R(D) \leq R_2(D),$$

where

$$R_2(D) = \frac{1}{2}\{1 + H(p) - 2H(D)\}, \quad o \leq D \leq D,$$

and

$$D_1 = \frac{1}{2}\{1 - \sqrt{1 - 2p}\}.$$

$H(p)$ is the entropy function as defined earlier. For rate $-\frac{1}{2}$ codes,

$$R_2(D) = \frac{1}{2}$$

and it follows that

$$H(D) = \frac{1}{2}H(p)$$

or

$$D_2 = H^{-1}\{\frac{1}{2}H(p)\}.$$

D_2 now serves as an upper bound on the distortion for the BSMS when the rate is one half. For reference, let D_o be the distortion obtained from the BMS with equiprobable digits. Note that D_2 depends on p, whereas D_o does not.

Equivalent Binary Symmetric Markov Sources

The BSMS can be generated in a simple manner using a binary memoryless source (BMS), a unit delay and a mod 2 adder. The circuit diagram is shown in Figure 2. The quantities X_n and Y_n are binary digits, where X_n is the output of the BMS and Y_n is the output of the BSMS at time t = n. They are related by the simple relationship

$$Y_n = Y_{n-1} \oplus X_n$$

and the BMS generates zeros with probability p, i.e., $P(X_n = o) = p$.

It may be recalled that a BMS with P(0) = p can be converted to a BMS with P(0) = 1-p by simply complementing all the output digits, or equivalently by the adding the all-ones sequence to the output. Because of this reversible transformation, these two sources are equivalent in the context of source encoding. It may be reasonable to wonder if there are similar equivalences between BSMS's.

Consider the BSMS shown in Figure 3 where the BMS generates zeros with probability (1-p), i.e., $P(Z_n = 0) = 1-p$. It turns out that if we add (mod 2) the alternating sequence

$$\{U\} = \ldots.01010101 \ldots..$$

to the output of Y_n' of Figure 3, then this BSMS is converted to that of Figure 2. The modified version of Figure 3 is shown in Figure 4.

Since the BSMS's in Figs. 2 and 3 are related by a reversible transformation they are equivalent in the sense of source encoding. We should add that a convolutional encoder may not necessarily encode equivalent BSMS's equally well. Indeed our simulations reveal that the distortion introduced may differ significantly when encoding equivalent sources. The sources are equivalent in the sense that a "sufficiently clever encoder" should be able to encode one as well as the other. However, we could always assist the convolutional encoder by trying various transformations which might result in lesser distortion.

Simulation Results

The BSMS was simulated for various values of p and then an exhaustive search was made for the convolutional encoders which generated the least distortion for a given constraint length. At this time, it was not considered feasible to search beyond constraint length $\nu = 7$. The simulation results are given in a number of tables and graphs at the back of this report.

In general, the simulations show that the distortion decreases as the constraint length increases for a given BSMS. However, there are a number of interesting exceptions. We recall, at this point, that a BSMS with parameter p is equivalent to a BSMS with parameter (1-p). Hence, it is worth comparing the data and curves for complementary values of p. For $p \geq 0.5$, we tend to notice a somewhat smooth decline in distortion as ν increases. However, for $p < 0.5$ the relationship becomes somewhat more erratic. We observe in particular that there is a code with $\nu = 3$ which performs especially well for a wide range of values of p, for $p < 0.5$. In fact, this code gives less distortion than any other code found for p in the range $0.05 \leq p \leq 0.25$. We are at present trying to determine why this code is such a good match for the BSMS over such a wide range of the parameter p.

Although exhaustive searches beyond $\nu = 7$ are currently not feasible, simulating the performance of a few selected codes of greater constraint length is quite reasonable. If the generators of the codes found are examined, a number of interesting patterns reveal themselves. It is believed that longer constraint length codes ($\nu > 7$) with these patterns should be simulated to see if the distortion continues to decline.

Finding the best code in each instance may be satisfying, but we may miss important classes of codes if we ignore all but the best. As a result, we often list the best two or three codes if their performances are comparable. To examine this issue more systematically, we have listed the distortions produced by all the codes of a given constraint length for a few values of p. This gives us the distribution of distortion for a given ν and a given BSMS. This data and the associated graphs are also given at the back of the report.

CONCLUSIONS

In this report, we have examined the ability of rate $-\frac{1}{2}$ binary convolutional codes to encode binary symmetric Markov sources (BSMS). For this scheme, the compression ratio if fixed at two hence the quantity of interest is the average distortion. As might be expected, the simulations show that distortion tends to decrease as the constraint length is increased. However, there are some striking exceptions to this trend. It was observed that for p in the range $0.05 \leq p \leq 0.25$, where p is the parameter of the BSMS, a code with $\nu = 3$ generated less distortion than any other code with constraint length $\nu \leq 7$. This code has generators $G_1(D) = D^2$ and $G_2(D) = 1 + D + D^2$. The theoretical basis for this is not understood and merits further investigation. For interest, we note that this code is systematic with free distance $d_F = 4$.

Several generator patterns emerged from the data and these are worth exploring for constraint length greater than seven.

REFERENCES

[1] T. Berger, "Rate Distortion Theory", Prentice-Hall, Englewood Cliffs, New Jersey, 1971, pages 49 and 253.

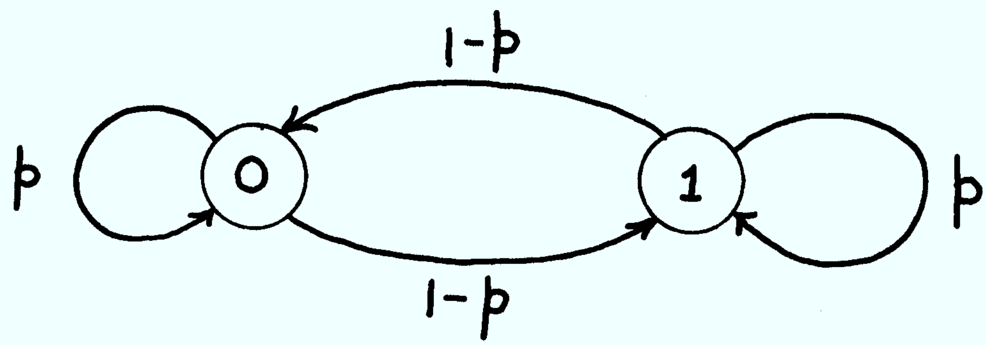

Fig. 1

STATE TRANSITION DIAGRAM FOR BSMS

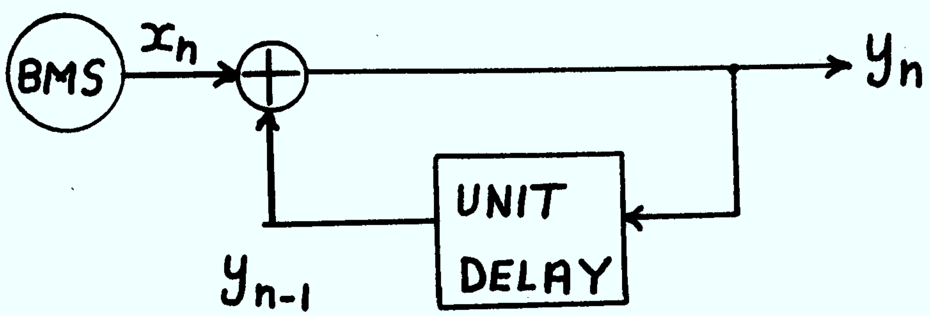

Fig. 2

CIRCUIT DIAGRAM FOR BSMS

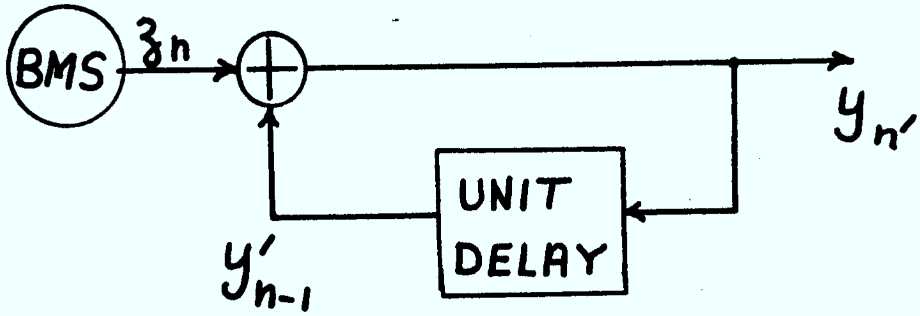

Fig. 3

BSMS WITH $P(z_n = o) = 1 - p$

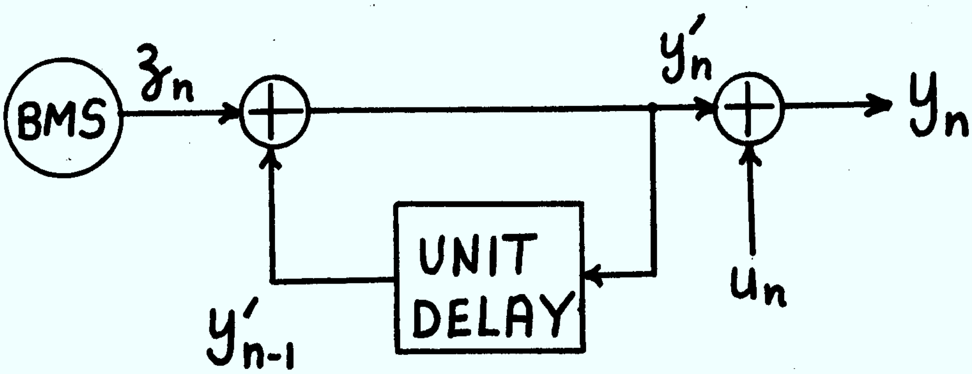

Fig. 4

MODIFIED VERSION OF FIG. 3 WHICH IS EQUIVALENT TO SOURCE IN FIG. 2.

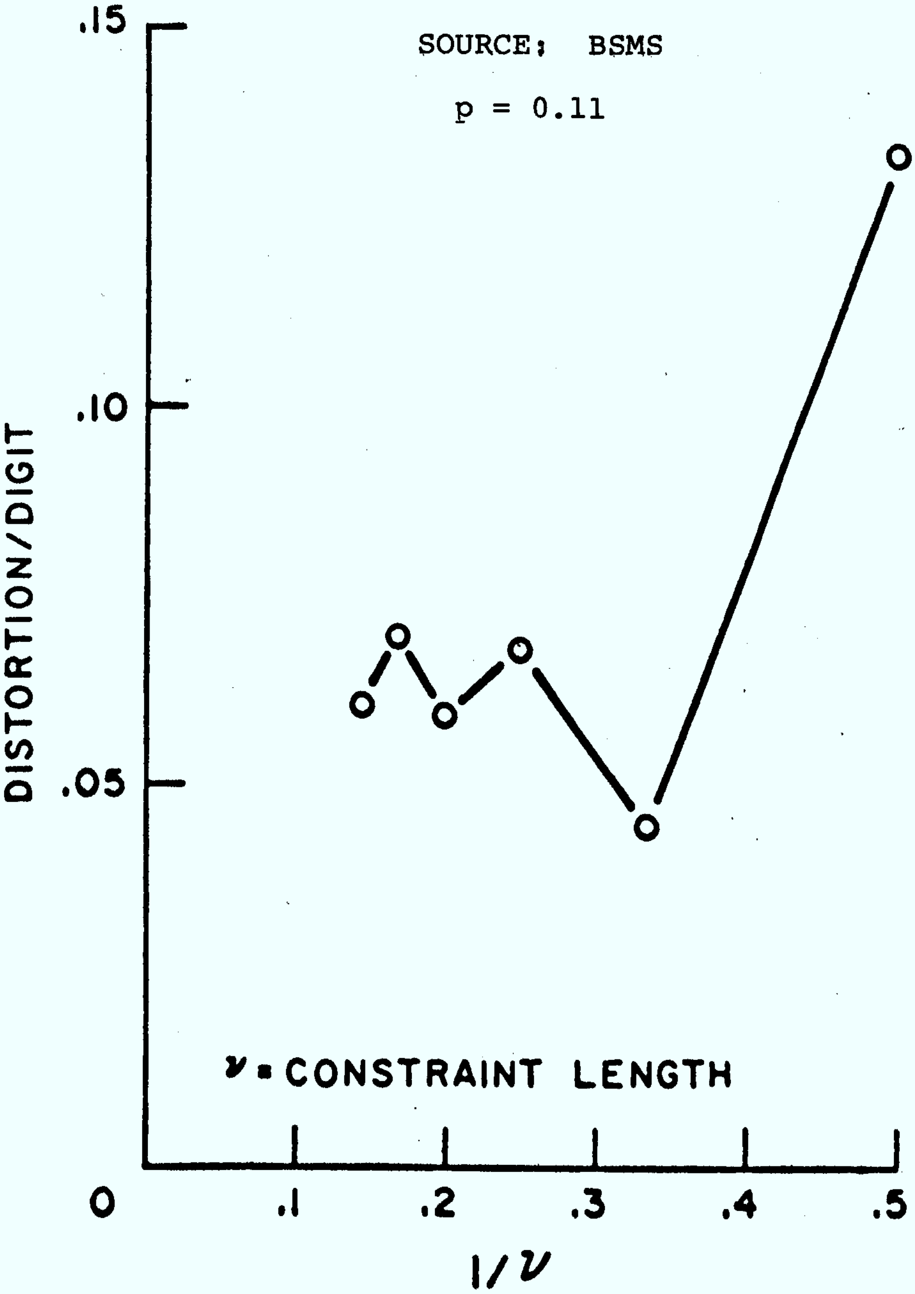

Fig. 1

DISTORTION VS RECIPROCAL OF CONSTRAINT LENGTH
FOR BINARY SYMMETRIC MARKOV SOURCE (BSMS) WITH
PARAMETER p = 0.11

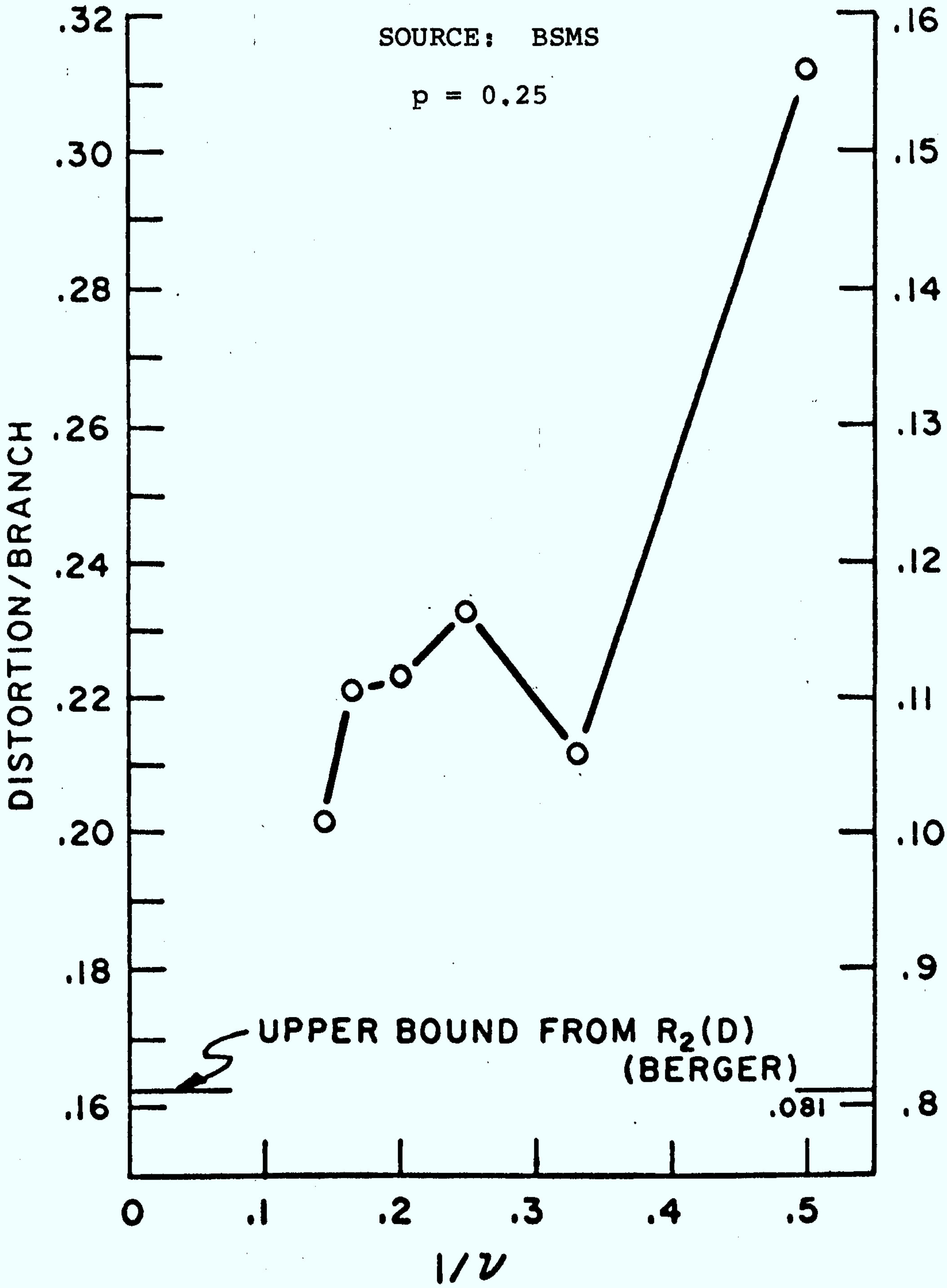

Fig. 2

DISTORTION VS RECIPROCAL OF CONSTRAINT LENGTH
FOR BSMS WITH PARAMETER p = 0.25

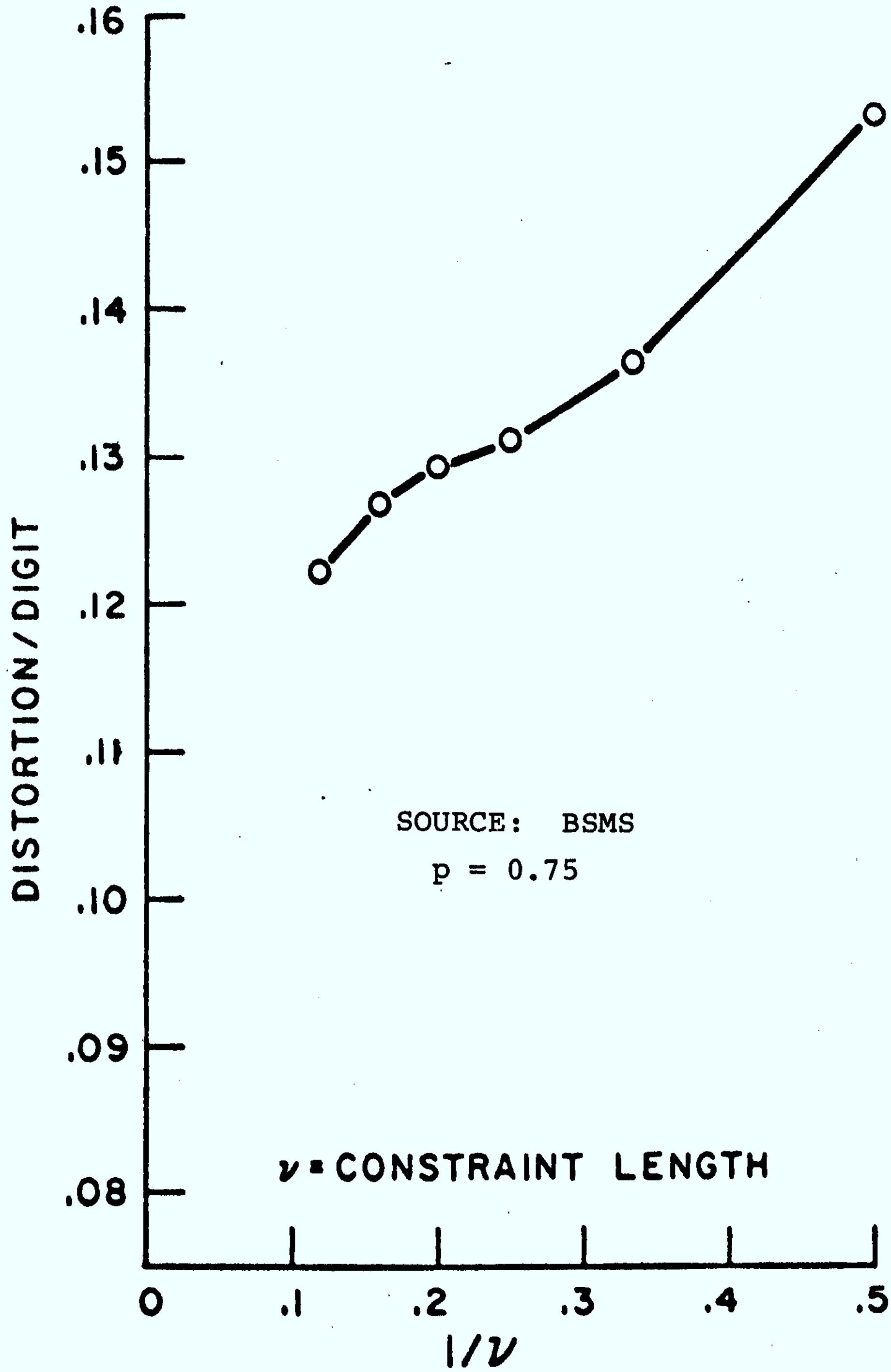

Fig. 3

DISTORTION VS RECIPROCAL OF CONSTRAINT LENGTH FOR BSMS WITH PARAMETER p = 0.75

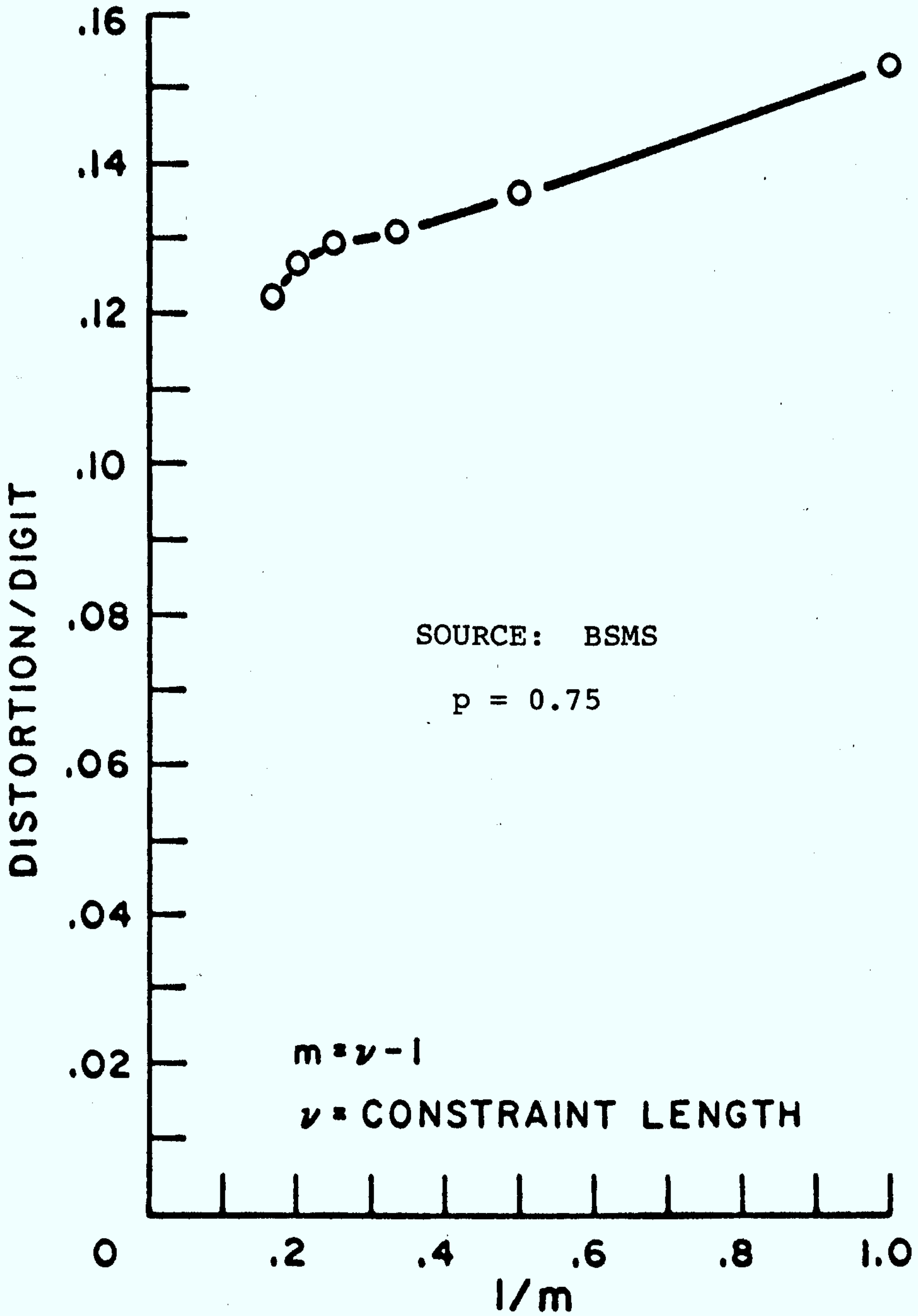

Fig. 4

DISTORTION VS RECIPROCAL OF m WHERE $m = \nu - 1$
FOR BSMS WITH PARAMETER p = 0.75

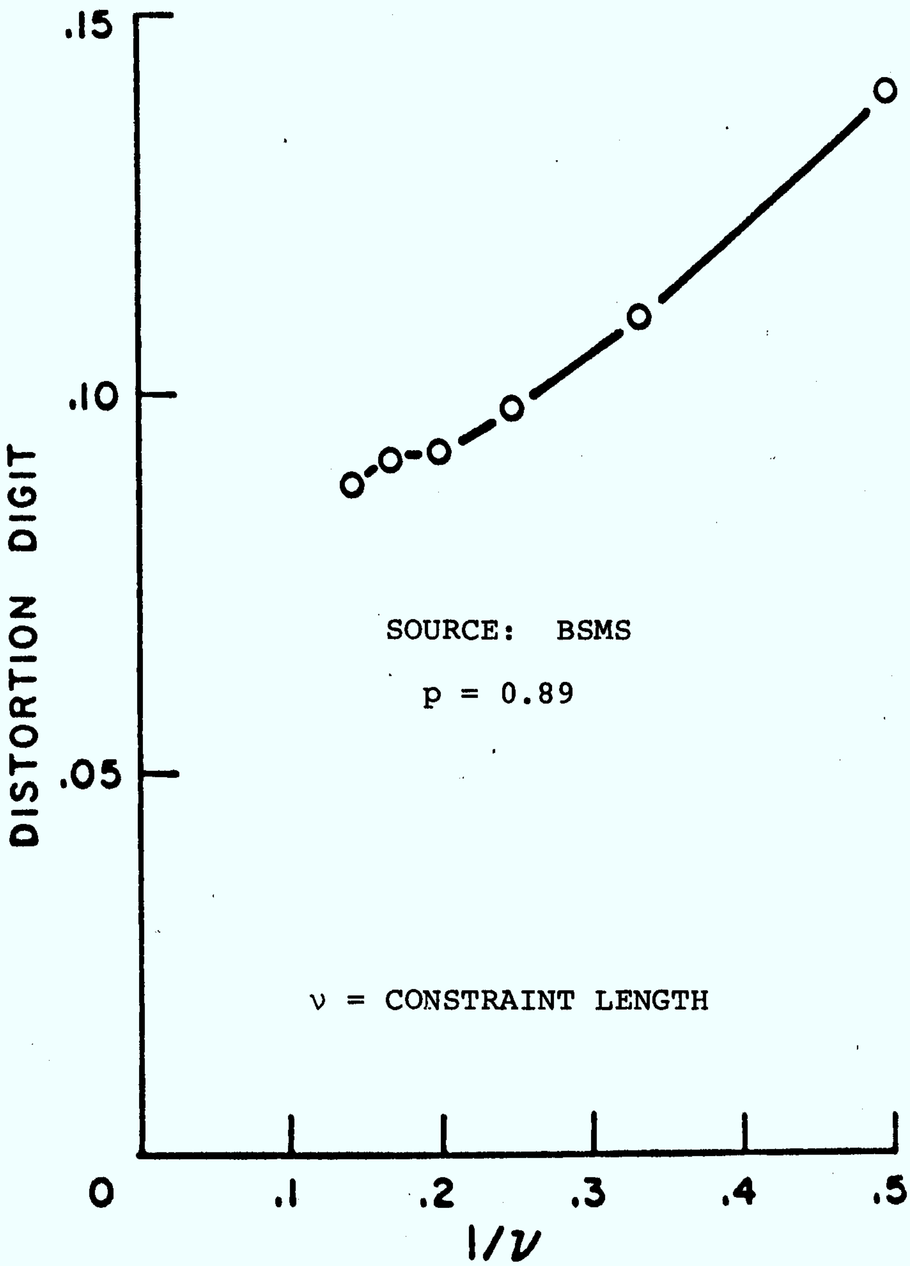

Fig. 5

DISTORTION VS RECIPROCAL OF CONSTRAINT LENGTH FOR BSMS WITH PARAMETER p = 0.89

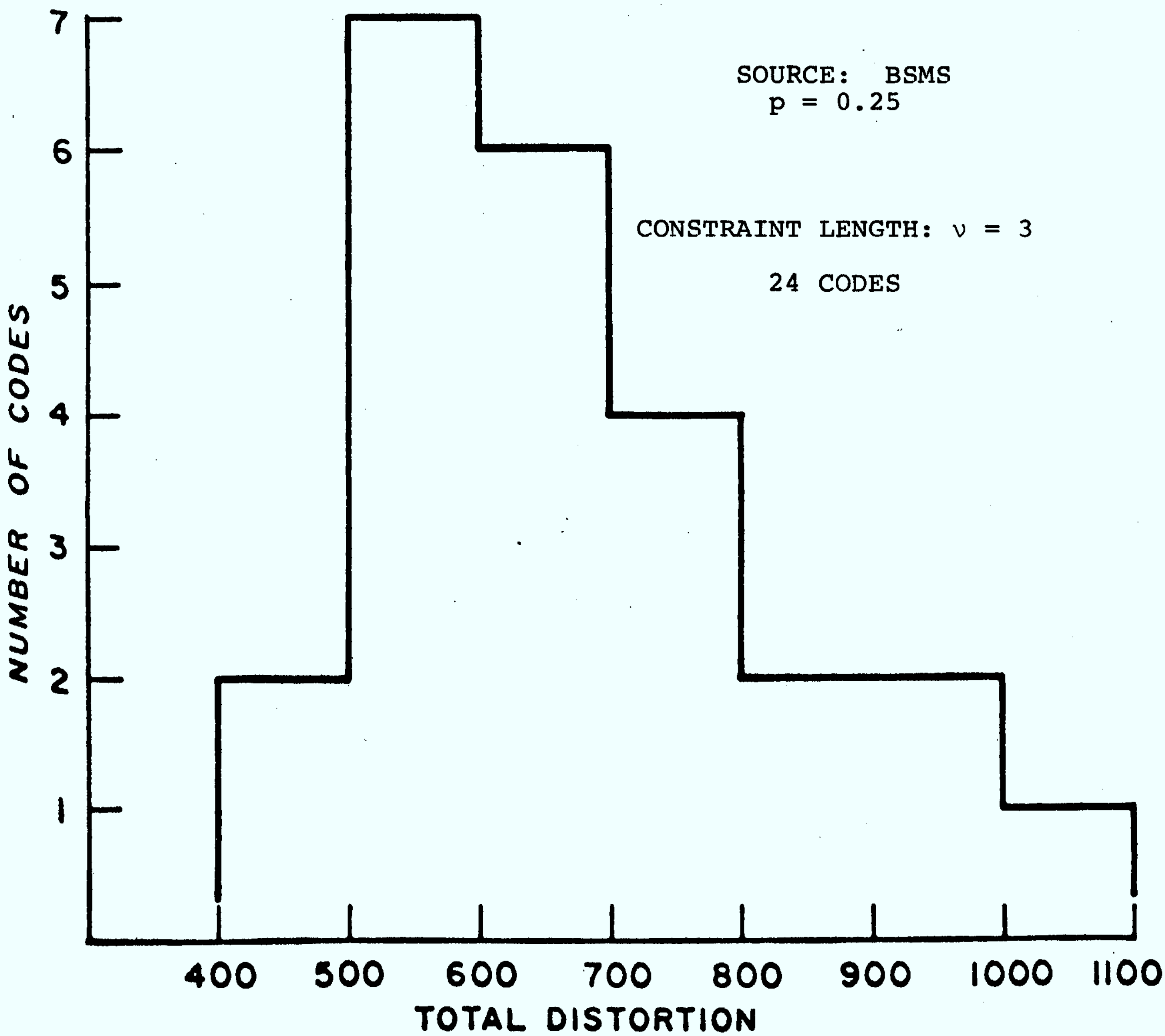

Fig. 6

DISTRIBUTION OF DISTORTION FOR BSMS WITH p = 0.25 USING CODES WITH CONSTRAINT LENGTH ν = 3.

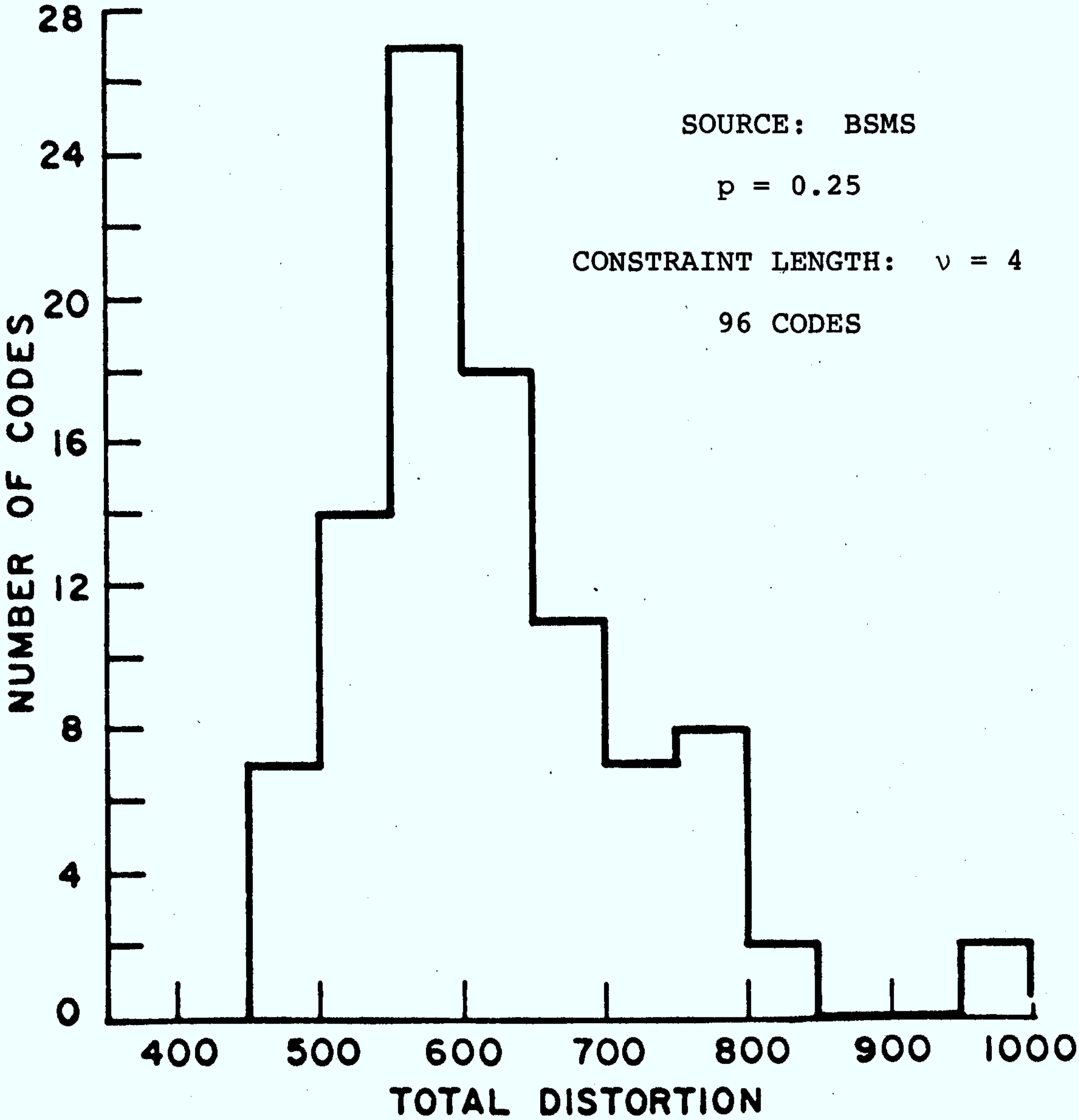

Fig. 7

DISTRIBUTION OF DISTORTION FOR BSMS WITH $p = 0.25$ USING CODES OF CONSTRAINT LENGTH $\nu = 4$.

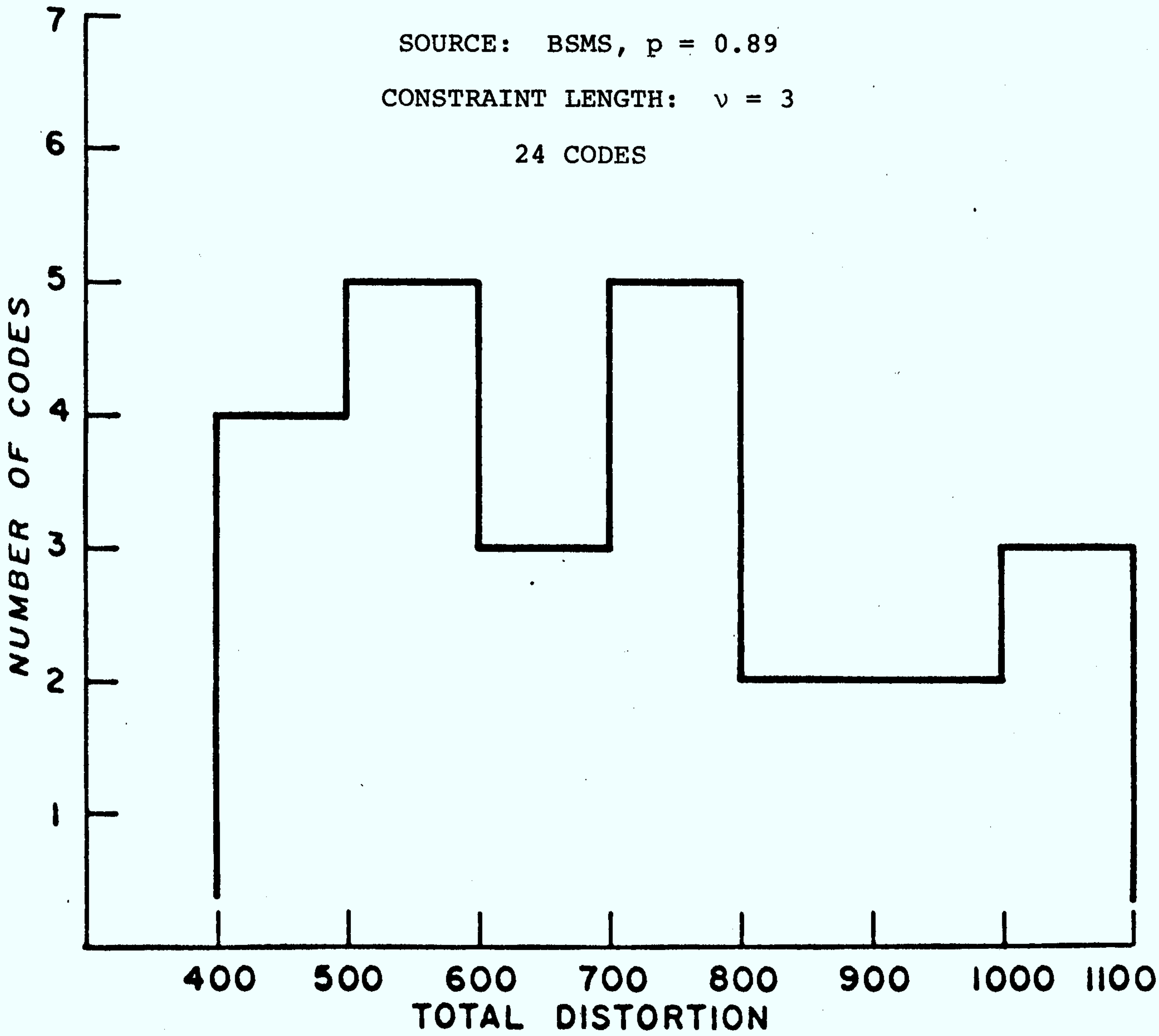

Fig. 8

DISTRIBUTION OF DISTORTION FOR BSMS WITH $p = 0.89$ USING CODES OF CONSTRAINT LENGTH $\nu = 3$.

DISTORTION vs CONSTRAINT LENGTH

SOURCE: BINARY MEMORYLESS SOURCE (BMS)

$p(0) = p(1) = 0.5$

ν	DIST./DIGIT	GENERATOR	GENERATOR POLYNOMIALS	FREE DISTANCE
2	.1672	11 01	$1+D$ D	3
3	.1371	101 111	$(1+D)^2$ $1+D+D^2$	5
4	.1350	1001 1011	$(1+D)(1+D+D^2)$ $1+D^2+D^3$	5
5	.1275	10001 10110	$(1+D)^4$ $1+D^2+D^3$	5
6	.1280	100001 111011	$(1+D)(1+D+D^2+D^3+D^4)$ $1+D+D^2+D^4+D^5$	7
	.1280	100001 111101	$(1+D)(1+D+D^2+D^3+D^4)$ $1+D+D^2+D^3+D^5$	7
7	.1210	1000001 1101110	$(1+D)^2(1+D+D^2)^2$ $1+D+D^3+D^4+D^5$	7

DISTORTION vs CONSTRAINT LENGTH

SOURCE: BSMS $p = 0.05$ 4000 DIGITS

ν	1/ν	TOT. DIST.	DIST./DIGIT	GENERATORS
2	.5000	503	.1258	11 10
		526	.1315	11 01
3	.3333	72	.0180	001 111
		110	.0275	011 100
		166	.0415	100 111
		191	.0478	110 001
4	.2500	140	.0350	1011 0100
		149	.0373	1101 0010
5	.2000	131	.0328	00001 11111
		132	.0330	01010 11111
		135	.0338	11111 11011
		138	.0345	01111 10000
6	.1667	166	.0415	001011 110111
		173	.0433	111011 000100
		175	.0438	101111 010000
		177	.0443	011010 100101
7	.1429	145	.0363	0000001 1111111

DISTORTION vs CONSTRAINT LENGTH

SOURCE: BSMS p = 0.11 4000 digits

ν	1/ν	TOT. DIST.	DIST./DIGIT	GENERATORS
2	.5000	532	.1330	11 10
3	.3333	179	.0448	001 111
4	.2500	272	.0680	1011 0100
5	.2000	237	.0593	11110 00111
6	.1667	279	.0698	001011 110111
7	.1429	243	.0608	0101010 1111111
		244	.0610	0011111 1111111

DISTORTION vs CONSTRAINT LENGTH

SOURCE: BSMS   p = 0.15   4000 digits

ν	1/ν	TOT. DIST.	DIST./DIGIT	GENERATORS
2	.5000	578	.1445	11 10
3	.3333	250	.0601	001 111
4	.2500	339	.0847	1011 0100
5	.2000	317	.0793	00001 10011
6	.1667	321	.0803	001011 110111

DISTORTION vs CONSTRAINT LENGTH

SOURCE: BSMS p = 0.25 4000 digits

ν	1/ν	TOT. DIST.	DIST./DIGIT	GENERATORS
2	.5000	624	.1560	10 01
3	.3333	423	.1058	001 111
4	.2500	466	.1165	1011 0100
		467	.1168	0001 1011
		468	.1170	0010 1101
5	.2000	446	.1115	00001 10011
6	.1667	442	.1105	000001 100011
7	.1429	403	.1008	0011001 1110011
		445	.1113	0000001 1000011

DISTORTION vs CONSTRAINT LENGTH

SOURCE: BSMS $p = 0.35$ 4000 digits

ν	$1/\nu$	TOT. DIST.	DIST./DIGIT	GENERATORS
2	.5000	645	.1612	11 10
3	.3333	546	.1365	101 111
4	.2500	538	.1345	1011 1111
5	.2000	505	.1262	00101 11111
6	.1667	485	.1213	011001 101010

DISTORTION vs CONSTRAINT LENGTH

SOURCE: BSMS $\quad$ $p = 0.65$ $\quad$ 4000 DIGITS

ν	1/ν	TOT. DIST.	DIST./DIGIT	GENERATORS
2	.5000	669	.1673	11 01
		669	.1673	11 10
3	.3333	555	.1388	101 111
4	.2500	549	.1373	1111 1011
		553	.1383	1001 1101
5	.2000	524	.1310	10001 11111
		531	.1328	10001 01101
6	.1667	509	.0848	100001 101111
		511	.1278	100001 110111
		512	.1280	100001 111101
		517	.1293	100001 111011

DISTORTION vs CONSTRAINT LENGTH

SOURCE: BSMS p = 0.75 4000 digits

ν	1/ν	TOT. DIST.	DIST./DIGIT	GENERATORS
2	.5000	611	.1528	11 10
		644	.1610	11 01
3	.3333	545	.1363	111 110
		553	.1383	101 111
		556	.1390	011 111
4	.2500	524	.1310	0111 1111
		527	.1318	1111 1110
		534	.1335	1111 1011
5	.2000	512	.1280	11111 11110
		517	.1293	10001 11111

SOURCE: BSMS p = 0.75 4000 digits (cont'd)

ν	1/ν	TOT. DIST.	DIST./DIGIT	GENERATORS
6	.1667	506	.1265	100001 101111
		506	.1265	100001 111101
		510	.1275	111111 111110
		511	.1278	100001 110111

DISTORTION vs CONSTRAINT LENGTH

SOURCE: BSMS $\hat{p} = 0.85$ 4000 DIGITS

ν	1/ν	TOT. DIST.	DIST./DIGIT	GENERATORS
2	.5000	582	.1455	11 10
		591	.1478	01 11
3	.3333	458	.1145	111 110
		477	.1193	011 111
4	.2500	439	.1100	1111 1110
		442	.1105	0111 1111
5	.2000	422	.1055	11111 11110
		436	.1090	01111 11111
6	.1667	420	.1050	111111 111110
		428	.1070	011111 111111

DISTORTION vs CONSTRAINT LENGTH

SOURCE: BSMS $p = 0.89$ 4000 digits

ν	$1/\nu$	TOT. DIST.	DIST./DIGIT	GENERATORS
2	.5000	557	.1393	11 10
3	.3333	439	.1098	111 110
4	.2500	392	.0980	1111 1110
5	.2000	369	.0923	11111 11110
6	.1667	365	.0913	111111 111110
7	.1429	353	.0883	1111111 1111110

DISTORTION vs CONSTRAINT LENGTH

SOURCE: BSMS   p = 0.95   4000 DIGITS

ν	1/ν	TOT. DIST.	DIST./DIGIT	GENERATORS
2	.5000	540	.1350	01 11
		555	.1388	11 10
3	.3333	389	.0973	011 111
		398	.0995	111 110
4	.2500	309	.0773	0111 1111
		322	.0805	1111 1110
5	.2000	272	.0680	01111 11111
		282	.0705	11111 11110
6	.1667	270	.0675	011111 111111
		270	.0675	111111 111110

DISTRIBUTION OF DISTORTION

SOURCE: BSMS, p = 0.25

LENGTH OF SIMULATION: 4000 DIGITS

$\nu = 3$, 24 CODES

DISTORTION GENERATED OVER 4000 DIGITS

423	528	626	706	857	918	1006
445	546	636	734	888	950	
	563	646	750			
	577	650	795			
	580	672				
	583	686				
	595					

RANGE OF DISTORTION	# OF CODES
400-499	2
500-599	7
600-699	6
700-799	4
800-899	2
900-999	2
1000-1099	1
TOTAL	24

DISTRIBUTION OF DISTORTION

SOURCE: BSMS, $p = 0.25$

LENGTH OF SIMULATION: 4000 DIGITS

CONSTRAINT LENGTH: $\nu = 4$

96 CODES

RANGE OF DISTORTION	# OF CODES
400-449	0
450-499	7
500-549	14
550-599	27
600-649	18
650-699	11
700-749	7
750-799	8
800-849	2
850-899	0
900-949	0
950-999	2
1000-1049	0
TOTAL	96

DISTRIBUTION OF DISTORTION

SOURCE: BSMS, $p = 0.75$

LENGTH OF SIMULATION: 2000 DIGITS

$\nu = 3$, 24 codes

DISTORTION GENERATED OVER 2000 DIGITS

274	338	394
274	345	397
274	369	416
292	376	418
320	377	424
326	382	485
328	384	487
334	389	516

RANGE OF DISTORTION	# OF CODES
250-299	4
300-349	6
350-399	8
400-449	3
450-499	2
500-549	1
TOTAL	24

DISTRIBUTION OF DISTORTION

SOURCE: BSMS, $p = 0.89$

LENGTH OF SIMULATION: 4000 DIGITS

$\nu = 3$, 24 CODES

DISTORTION GENERATED OVER 4000 DIGITS

439	509	617	768	803	938	1012
456	529	621	771	877	957	1046
496	593	677	783			1084
499	594		788			
	598		789			

RANGE OF DISTORTION	# OF CODES
400-499	4
500-599	5
600-699	3
700-799	5
800-899	2
900-999	2
1000-1099	3
TOTAL	24