#64

International Telegraph and Telephone
Consultative Committee
(CCITT)

Question 4/XV
Specialist group on coding
for visual telephony

English only
Date: December, 1985

STUDY GROUP XV - CONTRIBUTION NO

Source: UNITED STATES OF AMERICA

Title: A PROCEDURE FOR EVALUATING ALTERNATIVE VIDEO CODING TECHNIQUES

1.0 INTRODUCTION

The purpose of this contribution is to describe a methodology for the evaluation of motion television coding techniques for teleconferencing applications.

The principal characteristics of the coding techniques to be tested include the following:

- o Utilize digital communication channels
- o Provide color capability
- o Provide motion capability

The specific objective of the tests described is to rank the codecs tested in order of performance capability. The reason that new tests are being developed is that there are presently no agreed upon test procedures for that purpose. The tests described will utilize a specially prepared video tape containing still and motion sequences designed specifically for the evaluation of this type of coding technique. These sequences will be passed through

the codecs and the output recorded on video tape. The evaluation and grading of each codec will be on a subjective, comparative basis. The intent of CCIR Recommendation 500-2, Method for the subjective Assessment of the Quality of Television Pictures (Vol. XI, Part 1, XVth Plenery Assembly, Geneva 1982) will serve as a guideline.

2.0 TEST PHILOSOPHY

The specific objective of the recommended test procedure is to rank all of the candidate coding techniques as to relative performance; the philosophy proposed is as follows:

- Subjectively evaluate the performance of the coding techniques one with respect to the other, to determine which produces the best overall results,
- Generate a performance grade for each of the coding techniques relative to the best overall performance,

The tests consist of two basic parts:

- Gathering performance data on the candidate coding techniques
- Ranking the performance of the techniques tested.

The first step consists of feeding the video signal from the test video tape into the codec to be evaluated. The output video signal from the receive side of the codec is recorded on a video tape recorder without performing a grading evaluation at that time. The video tape recorders must be of high quality so that they will provide an excellent input video signal and will not affect the quality of the recorded output signal.

The second phase will determine a performance grade for each codec in comparison with the other codecs. The test consists of evaluating the performance of each codec as recorded on the video tape against each of the other codecs taken two at a time and

determining which performs better. The performance of each codec is ranked as much better, better, slightly better, or the same as, the performance of the codec against which it is being evaluated. Once all of the codecs have been ranked against each other, an overall grade can be developed for each. The best performing codec is determined as well as the ranking of the other codecs with respect to it.

3.0 TEST PROCEDURE

The first phase of the test program consists of passing specially designed video signals through the codec pair (transmitter and receiver) and recording the output which consists of the picture sequences recorded on video tape for subsequent subjective evaluation.

Figure 3-1 is a block diagram showing the test implementation. The signal source, a video tape recorder, is connected to the codec transmitter to be tested. The video signal from the video tape recorder is monitored for quality and level on a television picture monitor and on a television waveform monitor/vectorscope. The digital signal from the codec transmitter is directly connected to the codec receiver in a normal configuration. Only for the test which determines the effect of channel errors on codec performance, this connection is made through a bit error inserter. The receive codec is connected to a television waveform monitor/vectorscope, a television picture monitor, and a high quality video tape recorder. Note that level, impedance and termination criteria must be carefully observed.

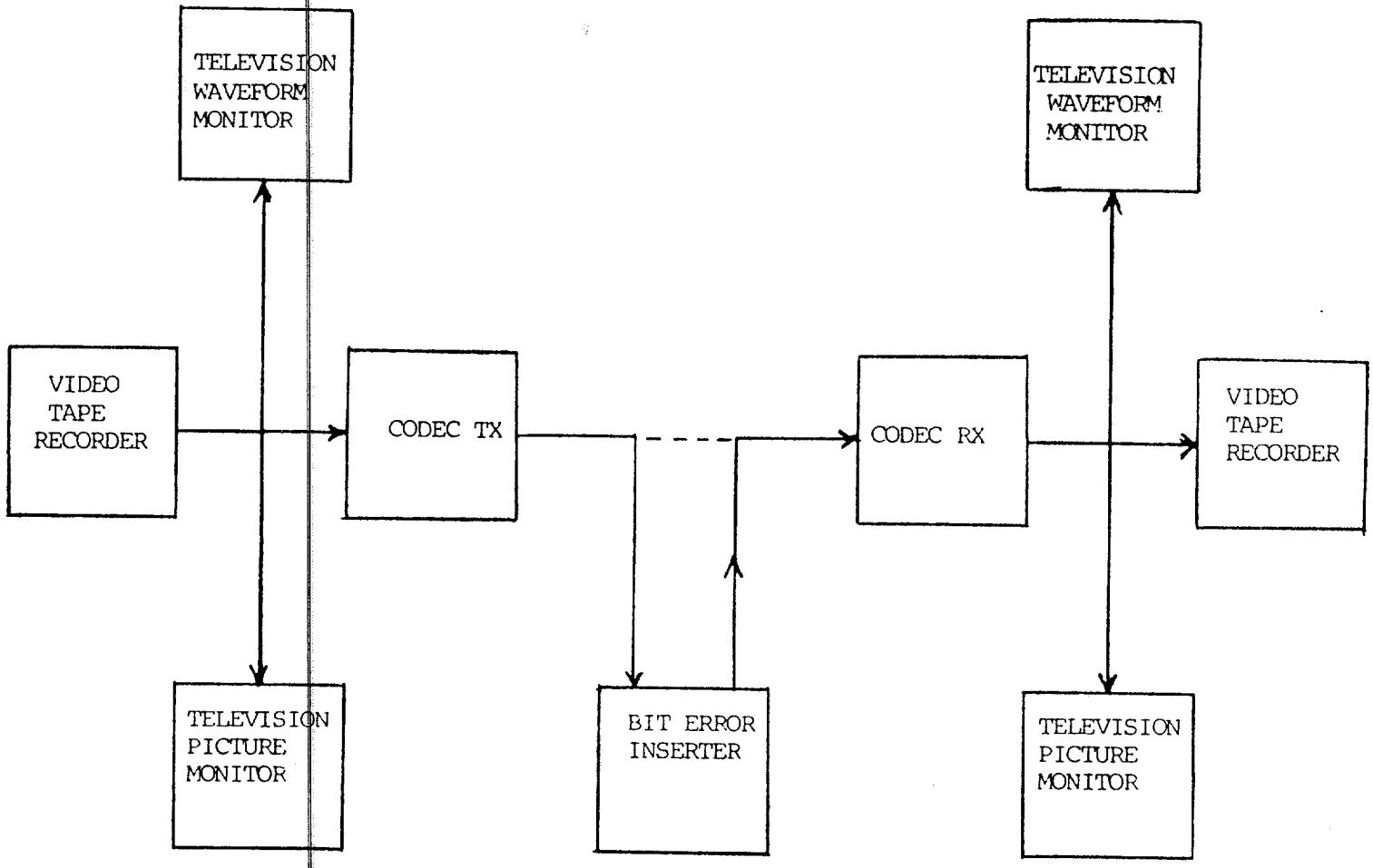


FIGURE 3-1 TEST IMPLEMENTATION AT THE MANUFACTURER"S FACILITY

4.0 EVALUATION PROCEDURE

The philosophy of evaluation is based on the fact that the category of video transmission equipments being evaluated are fairly new. Universally accepted objective tests to determine their performance have not yet been developed. Therefore, the performance must be evaluated subjectively.

The technique of comparison between codecs appears to provide the best method of grading the codecs, each with respect to all of the others. In this approach each codec is graded against each of the other codecs. Figure 4-1 shows the concept for a total of 5 codecs. Evaluating the 5 codecs against each other requires a total of 10 evaluation tests.

The grading scale shown on the codec evaluation form, Figure 4-2, is recommended. Since either of the codecs being evaluated can perform better than the other in any specific parameter, a scale which can rate either picture better than the other is necessary. The comparison scale of Figure 4-2 has this feature.

The basic concept of the evaluation procedure is very simple. It is shown in Figure 4-3. Two specifically prepared video tapes, each containing the same output pictures from different codecs in the same sequence, are displayed each on separate television monitors. The evaluators compare the quality of the two pictures and grade them on a comparative basis. The figure is deceivingly simple but each parameter of the test must be very carefully controlled to assure valid evaluation results. Since it is known that adjusting side-by-side color monitors for exactly equal

TEST	LEFT MONITOR CODEC NO.	RIGHT MONITOR CODEC NO.
1	1	2
2	1	3
3	1	4
4	1	5
5	2	3
6	2	4
7	2	5
8	3	4
9	3	5
10	4	5

TOTAL NUMBER OF TESTS REQUIRED = (N (N-1))/2

FIGURE 4-1 COMPARISON TESTS

CODEC EVALUATION FORM

EVALUATOR: _____ TEST NO: _____

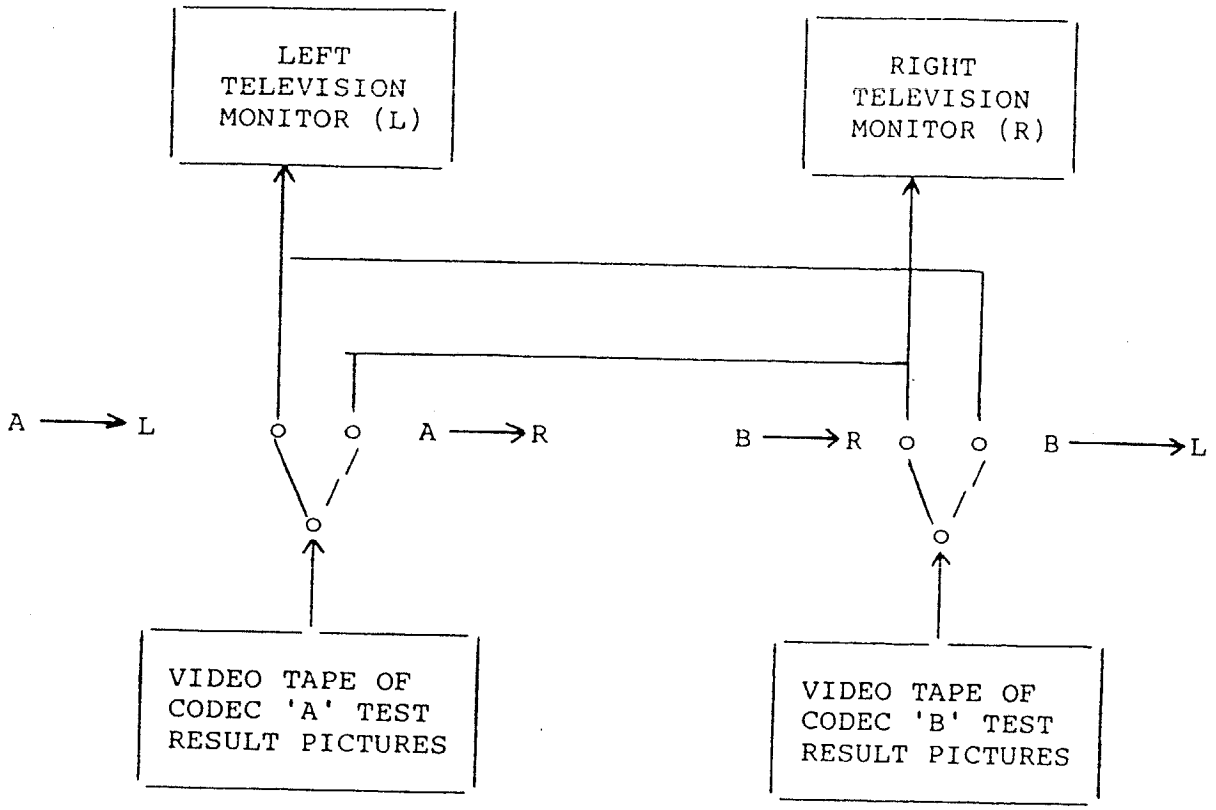
DATE: _____ PAGE: _____

+3; MUCH BETTER THAN	+2; BETTER THAN	+1; SLIGHTLY BETTER THAN	0; SAME AS

TEST SEQ.	LEFT PICTURE BETTER			SAME AS	RIGHT PICTURE BETTER		
1	+3	+2	+1	0	+1	+2	+3
2	+3	+2	+1	0	+1	+2	+3
3	+3	+2	+1	0	+1	+2	+3
4	+3	+2	+1	0	+1	+2	+3
5	+3	+2	+1	0	+1	+2	+3
6	+3	+2	+1	0	+1	+2	+3
7	+3	+2	+1	0	+1	+2	+3
8	+3	+2	+1	0	+1	+2	+3
9	+3	+2	+1	0	+1	+2	+3
10	+3	+2	+1	0	+1	+2	+3
11	+3	+2	+1	0	+1	+2	+3
12	+3	+2	+1	0	+1	+2	+3
13	+3	+2	+1	0	+1	+2	+3
14	+3	+2	+1	0	+1	+2	+3
15	+3	+2	+1	0	+1	+2	+3

TEST SEQ.	LEFT PICTURE BETTER			SAME AS	RIGHT PICTURE BETTER		
16	+3	+2	+1	0	+1	+2	+3
17	+3	+2	+1	0	+1	+2	+3
18	+3	+2	+1	0	+1	+2	+3
19	+3	+2	+1	0	+1	+2	+3
20	+3	+2	+1	0	+1	+2	+3
21	+3	+2	+1	0	+1	+2	+3
22	+3	+2	+1	0	+1	+2	+3
23	+3	+2	+1	0	+1	+2	+3
24	+3	+2	+1	0	+1	+2	+3
25	+3	+2	+1	0	+1	+2	+3
26	+3	+2	+1	0	+1	+2	+3
27	+3	+2	+1	0	+1	+2	+3
28	+3	+2	+1	0	+1	+2	+3
29	+3	+2	+1	0	+1	+2	+3
30	+3	+2	+1	0	+1	+2	+3

FIGURE 4-2 CODEC EVALUATION FORM



EQUIPMENT SETUP FOR EVALUATION OF CODEC PERFORMANCE, IN PAIRS. BY COMPARISON OF OUTPUT PICTURES.

FIGURE 4-3 EVALUATION SETUP

performance is very difficult, a monitor reversal switch is provided so that the effect of any small difference between monitors can be eliminated.

The characteristics of the viewing area must be controlled. Table 4-1 lists the more important recommended viewing conditions. The tabulation is an excerpt from CCIR REC 500-2.

One possible room layout is shown in Figure 4-4. The equipment is located on tables at both ends of the room. The monitor picture height was 15", therefore the ideal viewing distance is 7 1/2'. The front of each chair is located 7' from the center between the monitors which puts the eyes of the viewers at an average distance of about 8'. The variation in viewing angles should not make any noticeable difference at this distance. The overhead lighting is reduced to provide about 25 foot candles on the wall behind the monitors and about 15 foot candles at the chairs of the observers which gives just enough light for marking the score sheets. A low level work light on the equipment table is mainly used for threading the tape recorders. A screen prevents any light reflections on the monitor fronts and also serves to separate the operating personnel from the viewers.

The specific format of the video tape is shown in Figure 4-5. Each video tape consists of the same sequence of output pictures, each recorded through a different codec. The pictures are interspersed with neutral identification frames. The following describes the comparison procedure. Picture 'N' from codec 'A' is to be compared with picture 'N' from codec 'B'. The picture from

TABLE 4-1 RECOMMENDED SUBJECTIVE VIEWING CONDITIONS

PARAMETER	RECOMMENDATION (CCIR 500)
RATIO OF VIEWING DISTANCE TO PICTURE HEIGHT	4 TO 6
PEAK SCREEN LUMINANCE	70 CD/SQ. M
RATIO OF INACTIVE SCREEN TO PEAK LIMINANCE	<0.02
RATIO OF BACKGROUND LUMINANCE TO PEAK SCREEN LUMINANCE	0.15
AMBIENT ILLUMINANCE	LOW
CHROMATICITY OF SURROUND	D65

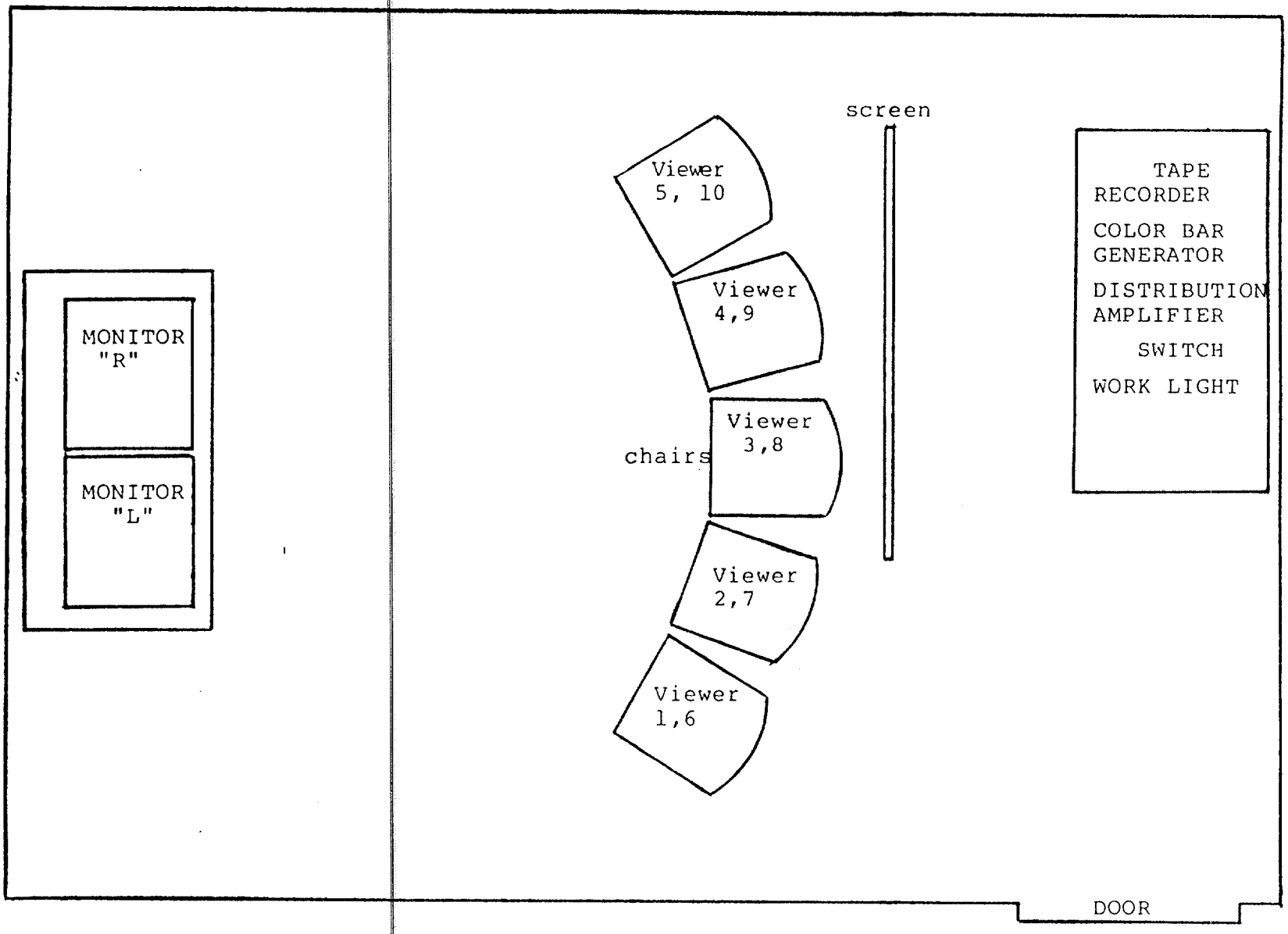


FIGURE 4-4 COMPARISON TEST ROOM LAYOUT

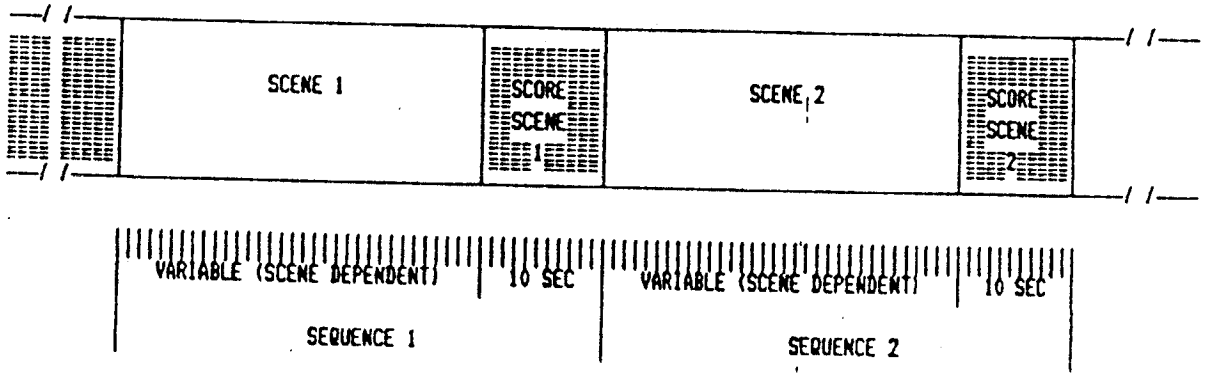


FIGURE 4-5 STIMULUS PRESENTATION

codec 'A' is presented on the left hand monitor: The picture from codec 'B' is presented on the right hand monitor. They are displayed for a period of time to permit adequate comparison, not less than about 15 seconds, followed by a 10 second display of a neutral field containing the caption "Score Sequence N". This will permit the evaluators adequate time to record the grade they have assigned to the better performing codec. The process is repeated for all pictures from 1 to N as shown in Figure 4-1.

The switch which is provided to reverse the physical order of the displays permits showing a duplicate of a previously presented picture on the opposite monitor; that is, the picture previously displayed on the right hand monitor now appears on the left and vice versa. This is shown for two selected sequences in Figure 4-6. These sequences are also graded by the evaluators. Correlating the results of the reversed display with the original will provide an additional degree of confidence in the tests. Display reversal for half of the test tape is also possible. The total test sequence should be limited to about 30 minutes.

Figure 4-2 shows the suggested codec evaluation form on which the evaluators record the grade which they assign to each sequence. It contains headers and spaces for all of the pertinent data to define the test run and the evaluator. The grading scale is printed on this page to serve as a reference should the evaluator need it. The recording procedure is to place a mark in the box containing the appropriate grade for each sequence. This format assures consistent recording of grades with an absolute minimum of distraction for the evaluator.

SEQUENCE	PICTURE	CODEC	MONITOR
1	1 1	A B	LEFT RIGHT
2	2 2	A B	LEFT RIGHT
≈	≈	≈	≈
(N-1)	(N-1) (N-1)	A B	LEFT RIGHT
(N)	(N) (N)	A B	LEFT RIGHT
≈	≈	≈	≈
CHECK 1	SELECT 1 SELECT 1	A B	RIGHT LEFT
CHECK 2	SELECT 2 SELECT 2	A B	RIGHT LEFT

THIS SEQUENCE PERMITS THE COMPARATIVE EVALUATION (GRADING) OF THE RELATIVE PERFORMANCE OF TWO CODECS. EACH PICTURE ON THE TEST TAPE (1 TO N) HAS BEEN TRANSMITTED THROUGH EACH CODEC, A AND B, THE OUTPUT PICTURE FROM CODEC A IS PRESENTED ON THE LEFT MONITOR: THE OUTPUT PICTURE FROM CODEC B IS PRESENTED ON THE RIGHT MONITOR. THE EVALUATORS SELECT THE HIGHER QUALITY PICTURE AND GRADE IT ON A COMPARATIVE BASIS. SEVERAL CHECK SEQUENCES ARE PROVIDED (TWO ARE SHOWN IN THE FIGURE ABOVE). IN THESE SEQUENCES PICTURES ALREADY PRESENTED ARE SHOWN BUT WITH THE CODEC/MONITOR RELATIONSHIP REVERSED. THIS WILL PROVIDE AN ADDED DEGREE OF CONFIDENCE IN THE TEST RESULTS. THESE CHECK SEQUENCES MAY BE INTERSPERSED WITH OTHERS AND MONITOR REVERSAL FOR HALF OF THE TAPE MAY BE USED.

FIGURE 4-6 EVALUATION SEQUENCES

Evaluation Computation

The preceding sections described the method of generating data from which a quantitative evaluation of codec performance can be ascertained. The following is a description of the calculations which produce a single quantitative grade for a codec's performance as compared to the performance of similar codecs.

The concept of comparing codecs A and B is shown in Figure 4-7. The major matrix in this figure is a planar matrix which lists the sequences evaluated along the ordinate and the evaluators along the abscissa.

The first calculation determines a mean and a standard deviation for each sequence as indicated by the arrows. The mean indicates the comparative performance of the codecs for each sequence. Since the reaction of each codec to different sequences is completely variable, the mean values may cover the whole range from +3 to -3 and be entirely valid. However, a high standard deviation indicates wide disagreement between evaluators. Should this occur for a specific sequence in several codec comparisons, it shows that scoring of this sequence is unduly difficult and it may be advisable to exclude it from the evaluation.

The second calculation is to determine the mean and the standard deviation for each of the evaluators. The rationale is similar to before: specifically, a mean far out of line with the means of the other evaluators provides some concern as to the

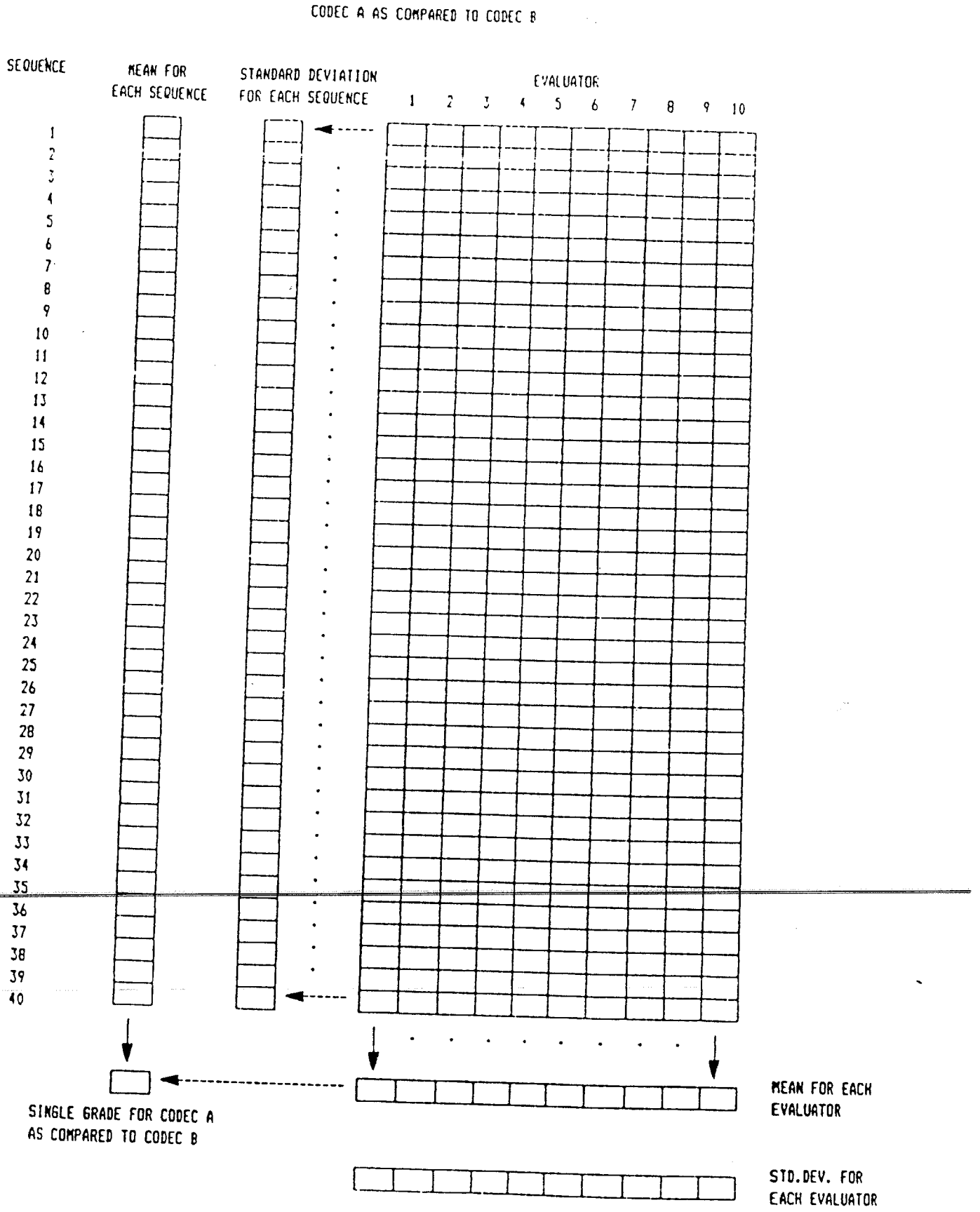


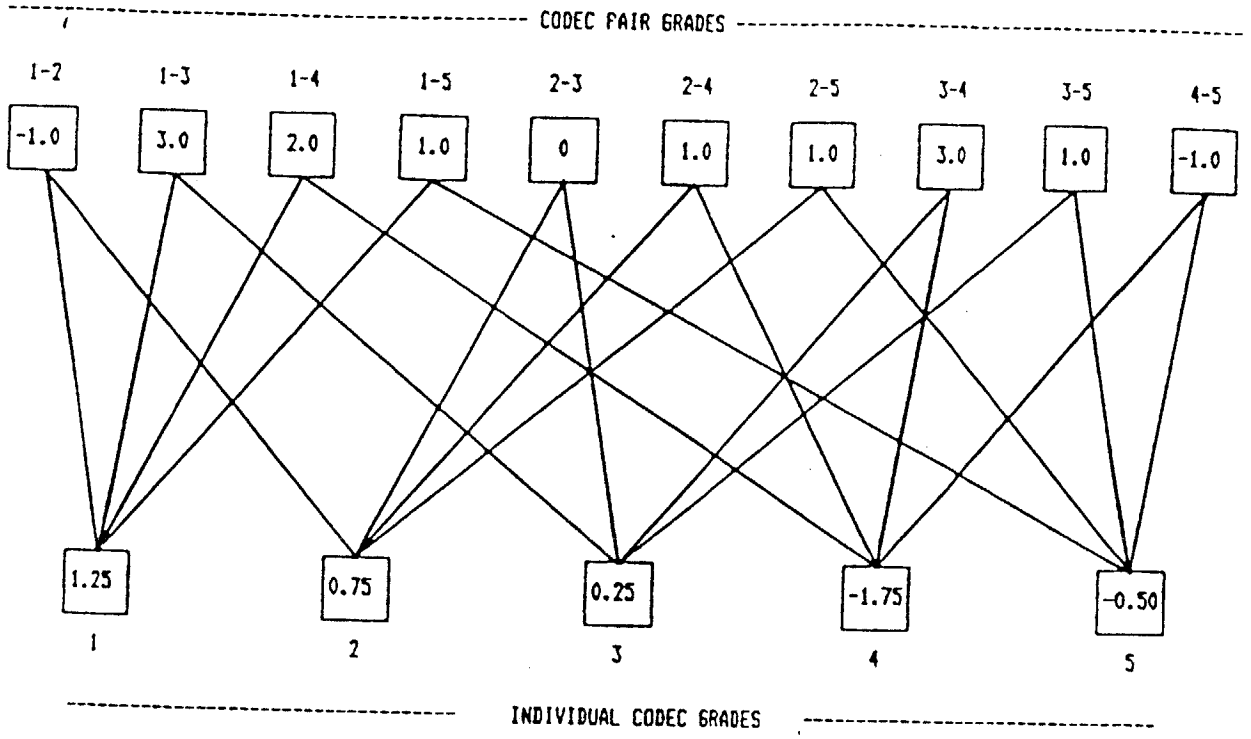
FIGURE 4-7 CODEC A AS COMPARED TO CODEC B

meaningfulness of the scoring by that particular evaluator, particularly if it recurs in several codec comparisons. The standard deviation, however, is determined by the range of scores used by each evaluator. Therefore, a high standard deviation mainly indicates a very outspoken and determined evaluator and by itself is no reason to question the validity of the result.

It remains then to produce a single grade for that specific codec comparison; eg., codec A compared to codec B. That single grade is determined as the mean of the means of the individual evaluators or of all the sequences employed. Both calculations cover all evaluators and test sequences and must yield the same grade. If this grade is positive, codec A has performed better than codec B: If this grade is negative, codec B has performed better than codec A. In this manner a single grade can be developed for each codec comparison.

The next set of calculations will rank the codecs. This is depicted graphically in Figure 4-8. To do this, a single grade must be developed to indicate the performance of each codec as compared to all other codecs. If, for example, there are 5 codecs, a single grade must be developed for the performance of codec A as compared to codec B, C, D, and E. In the previous paragraph the method of determining a grade for the evaluation of codec A as compared to codec B was presented. By extension, this same technique is employed to determine a single grade for the performance of codec A as compared to codec C, A to D, and A to E. It follows then that the single grade for the performance of codec A as compared to all other codecs is the mean of these individual

EVALUATION CALCULATION CONCEPT



RANK	CODEC	GRADE
1	1	1.25
2	2	0.75
3	3	0.25
4	5	-0.50
5	4	-1.75

FIGURE 4-8 EVALUATION CALCULATION CONCEPT

grades. This same procedure applies to determining a single performance grade for codecs B, C, D, and E. Note that the grade of codec B as compared to codec A is the negative of the grade for codec A as compared to codec B. Figure 4-8 shows the 10 scores of individual codec pairs on top. The single grades for each codec are shown in the 5 boxes below, with the connecting lines indicating how each of the single grades was derived as the mean of 4 individual scores. The ranking procedure is now simply a case of ranking the overall performance grades of the individual codecs.