COMMUNICATIONS
RESEARCH CENTER

CENTRE DE
RECHERCHES SUR
LES COMMUNICATIONS

MUSICAM

Listening Tests Report

Communications
Canada

Canada

CRC-RP-91.001

MUSICAM

Listening Tests Report

by

Ted Grusec, Ph.D.
Behavioral Research

and

Louis Thibault
Broadcast Systems and Networks Research

Communications Research Centre
Ottawa, Ontario, Canada

Approved for issue
as a CRC Report by:

William Sawchuk, DGBT

Table of Contents

List of Figures

List of Tables

Executive Summary

This report describes the content and the results of a series of listening tests that were carried out at the Communications Research Centre, Ottawa, Canada on the MUSICAM audio source coding system. The particular MUSICAM system that was tested was the version submitted to the ISO-IEC/MPEG committee: it performed independent coding of left and right channels of a stereo pair at a reduced bit-rate of 128 kbits/s per monophonic channel. An error protection scheme, which protected the most important coded information and which occupied about 10% of the compressed bit stream, was also implemented. The tests were carried out from November 1990 to January 1991 with an additional experiment run at the end of April 1991.

The following tests were performed:

1) Basic Audio Quality
2) Stereophonic Image Quality
3) Monophonic Compatibility
4) Robustness to Bit Errors
5) Tandem Coding Capability
6) FM vs MUSICAM Comparison

In four of these experiments (1, 3, 4 and 5), A-B presentation was used in conjunction with a continuous version of the 5-grade impairment scale described in CCIR Rec. 562-2. A continuous version of the 7-grade comparison scale described in Rec. 562-2 with A-B presentation was used in the FM vs MUSICAM comparisons. The Stereophonic Image Quality experiment used a diagram where listeners indicated the perceived spatial location of binaural auditory stimuli that were recorded with a dummy-head and reproduced over headphones.

A number of listeners ranging from 19 to 35 took part in the various experiments described above. Strictly speaking, few in the listening panel would qualify as "expert" listeners in the sense of having extensive experience in critical listening to potential digital coding distortions. The majority of these listeners could however be qualified as "skilled" or "experienced" because of their professional background or their special interest in audio. All listeners were administered the Seashore Tests of Musical Talent. This test attempts to assess an individual's music related profile in six categories: pitch, loudness, rhythm, time, timbre and tonal memory. Our interest in these measures is as a device to "calibrate" listeners as judges of music-related materials such as was used in many of the experiments reported here. Even listeners who were categorized as "low" on the Seashore tests were mostly above average in general population terms.

Two different experiments were conducted to assess the Basic Audio Quality of the MUSICAM system. The two were quite similar and differed mainly in that low anchor stimuli (i.e. deliberately impaired sequences) were used in the first experiment and not in the second one. In both experiments, listeners were unable to detect any significant differences between reference and MUSICAM encoded-decoded audio materials. Based on the listeners and experimental procedure used, the MUSICAM system tested appears to be transparent with respect to Basic Audio Quality.

In the Stereophonic Image Quality test, the binaurally recorded stimuli used produced a great deal of "mirroring" where events recorded in front of the dummy-head were often perceived at the rear by most listeners. Mirror transformations were used to reveal the systematic and symmetrical relationship between the objective and subjective localizations. These transformations did not obscure the comparisons between reference materials and MUSICAM processed ones: both were perceived identically. The MUSICAM system was thus found to be transparent with respect to the Stereophonic Image.

In the Monophonic Compatibility test, stereophonic audio materials processed through the MUSICAM system were graded identically when presented in mono or in stereo. Besides the intrinsic loss of the stereo image, the mixing of left and right independently coded channels did not produce additional impairment. And so, the MUSICAM system appears to be compatible with monophonic reproduction. The only caution to this conclusion is that absence of any difference between the stereo and mono modes may have been due to other factors such as the low criticality of the test sequences used. We temper our conclusion accordingly. Although the experiment supports monophonic compatibility for MUSICAM, additional testing is needed before a stronger conclusion can be made.

In the Robustness to Bit Errors experiment, random errors with gaussian-like distribution were injected in the MUSICAM compressed bit stream at rates ranging from $5x10^{-5}$ to $5x10^{-3}$. Error rates as low as $5x10^{-5}$ were found to produce "slightly annoying" audible degradation on one audio material (Glockenspiel). Error rates of $1x10^{-3}$ or more were necessary to produce a "slightly annoying" impairment to the other two audio materials used in the experiment. This conclusion however is only valid for the particular error protection scheme implemented in the MUSICAM system version tested.

The Tandem Coding experiment investigated the subjective quality of audio materials processed through 1 to 4 coding stages at 192 kbits/s followed by 2 or 5 coding stages at 128 kbits/s. No conversion to analog was done between coding stages. The combination of 1 to 4 stages at 192 kbits/s with 2 stages at 128 kbits/s was found to be transparent. From this it can be deducted that up to 4 stages at 192 kbits/s alone or 2 stages at 128 kbits/s alone should also be transparent. A cascade of 5 coding stages at 128 kbits/s was found to generate a "slightly annoying" impairment on one audio material (Glockenspiel) and a "perceptible but not annoying" impairment on the other two audio materials we used. The experiment did not explore cases of 3 or 4 stages at 128 kbits/s.

In the FM vs MUSICAM comparison, audio materials processed through MUSICAM (at 128 kbit/s per monophonic channel) were reliably preferred to FM although by a very small margin. The high quality FM signals used in the comparison were generated under ideal conditions which are not representative of typical FM reception by consumers.

The evidence for the usefulness of music judgement tests, such as the Seashore, in experiment of this type, is minor. Interesting but small differences between "high" and "low" scoring listeners were found only on the Basic Audio Quality experiments and in the FM vs MUSICAM comparison. "High" scorers appeared to be more critical listeners but none of the conclusions would be altered if the Seashore data was excluded as a factor in the analysis. As noted above, our listeners represented only a narrow, upper range in music judgement and so, in comparison to the general population, most were above average.

1. Introduction

A series of DR (Digital Radio) field trials and demonstrations were conducted in four major Canadian cities during the summer of 1990. The technologies used for this project included the MUSICAM audio source coding and the COFDM channel coding systems. The purpose of these field trials and demonstrations was to evaluate the performance of these technologies in the Canadian broadcast context and to increase the awareness of the Canadian broadcast community about Digital Radio.

This project was sponsored jointly by the CAB (Canadian Association of Broadcasters), the CBC (Canadian Broadcast Corporation), the DOC (Department of Communications) and the CRC (Communications Research Centre), all under the umbrella of the late CABSC (Canadian Advanced Broadcast Systems Committee). These field trials also included the partnership of the CCETT (Centre Commun d'Etudes de Télédiffusion et Télécommunications) in Rennes, France and the IRT (Institut Für Rundfunktechnik) in Munich, Germany which developed the above mentioned. technologies and provided the equipment and some technical support for the tests.

In the context of the Digital Radio evaluation program, a series of listening tests were carried out at the CRC from November 1990 to January 1991 with an additional test run at the end of April 1991. The main objectives sought in conducting these listening tests were the following:

a) to evaluate the performance of the MUSICAM system with respect to:

- basic audio quality
- stereophonic image quality
- monophonic compatibility
- robustness to bit errors
- tandem coding capability

b) to compare the basic audio quality of the MUSICAM system to that delivered by high-quality FM

c) to give the Canadian broadcasters an opportunity to assess the capability of this low bit-rate coding system in controlled listening conditions.

The purpose of this document is to describe the test procedures, equipment set-up and results of the listening tests. The particular MUSICAM source coding system that was tested is the version that was submitted in July 1990 to the ISO-IEC/MPEG (International Standards Organisation-International Electrotechnical Commission/Moving Picture Experts Group) operating at a compressed bit-rate of 128 kbits/sec per monophonic channel.

2. Facilities and Data Analysis

2.1 Hardware

2.1.1 MUSICAM system

The MUSICAM hardware used for the listening test consisted in an encoder supplied by the IRT and a decoder supplied by the CCETT. MUSICAM[1,2] is a perceptual audio coding system that decomposes the incoming PCM encoded time domain signal into 32 equal bandwidth subbands in the frequency domain. The coefficients at the output of each subband are adaptively quantized according to the masking properties of the human auditory system. The quantized subband coefficients are transmitted to the decoder which reconstructs the PCM time domain signal by means of a 32 subband synthesis filter bank.

The unit tested operated at a compressed bit-rate of 128 kbits/s per monophonic channel and the coding algorithm was the version submitted to the ISO-IEC/MPEG for its Stockholm listening tests. The left and right channels of the stereo pair were encoded and decoded independently. Input to the encoder and output from the decoder were done via the AES/EBU digital interface at a sampling rate of 48 kHz and 16 bits/sample.

An error protection scheme was also implemented to protect the most important coded information. A Golay (24,12) block code that could correct any 3 or fewer errors and detect 1 error within a 12 bit information word was used to protect the following information: the bit allocation and the scale-factor selection information, the two MSB (Most-Significant-Bits) of the samples of the first subband and the MSB of the samples of the second and third subbands. A parity bit was applied to the 4 MSB of each 6-bit scale-factor. An error occurring within these 4 MSB could be detected and concealed. The overall capacity required by this protection scheme was about 13 kbits/s, that is about 10% of the coded bit stream of 128 kbits/s.

2.1.2 Sampling rate conversion

Most of the sequences used in the listening tests were extracted from compact disks. The sampling rate conversion from 44.1 to 48 kHz was done using a Sony DFX2400 Sampling Rate Converter according to the arrangement shown in Figure 1 below.

2.1.3 Generation of MUSICAM material

The MUSICAM material was generated with the equipment configuration shown in Figure 2 below. Reference test sequences were recorded on a master source tape and fed to the input of the MUSICAM encoder by means of a source DAT recorder and the AES/EBU interface. The MUSICAM encoder and decoder were connected "back-to-back" through a custom interface. The decoded sequences were recorded with a target DAT recorder. A single pass through the encoder and decoder was done to produce the material used in all tests except the Tandem Coding.

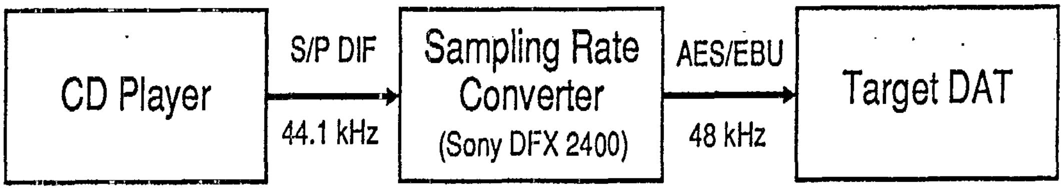


Figure 1 Equipment configuration for sampling rate conversion

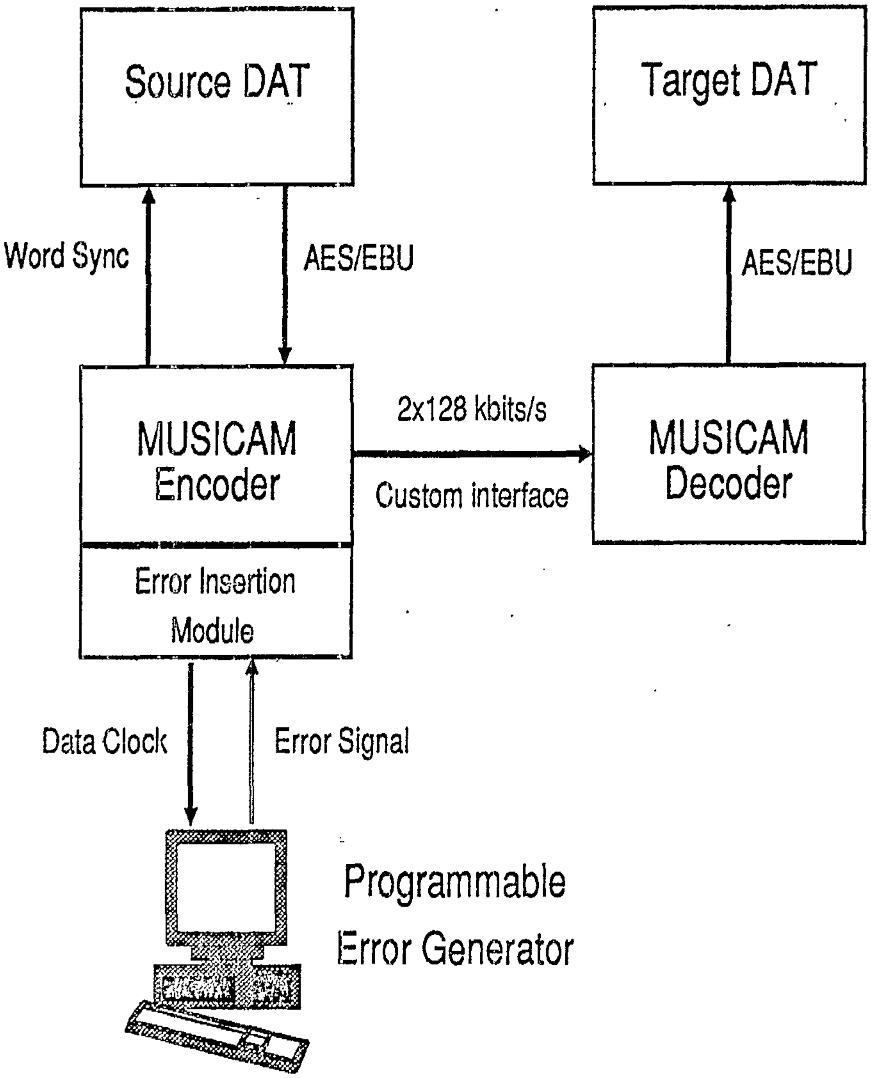


Figure 2 Equipment configuration for generating MUSICAM material

Material for the Tandem Coding test was produced by MPR Teltech in Vancouver, B.C which was carrying out a second listening tests program[3] focused on a version of MUSICAM operating at 192 kbits/s. This MUSICAM equipment was provided by the IRT and could be configured to operate at a compressed bit-rate of either 128 or 192 kbits/s per monophonic channel. An equipment configuration similar to that of Figure 2 (except for the Programmable Error Generator) was used. The reference material was processed through the MUSICAM equipment and recorded on a target DAT cassette to produce generation 1. This DAT cassette was then loaded into the source DAT recorder and processed again through the MUSICAM equipment to produce generation 2. This procedure was repeated up to the desired number of coding stages in tandem. No conversion to analog was involved in this process.

The test material to assess the Robustness to Bit Errors was produced using the Programmable Error Generator and the equipment configuration shown in Figure 2. This Error Generator was developed at the CRC and used a general purpose counter-timer board inserted in a PC-AT type personal computer. The board was controlled by a FORTRAN software that also calculated the random time interval between error events. Each error event consisted in a single bit error. The error rate was specified by the user. The random time interval between error corresponded to the sum of two uniformly distributed random variables: the resulting distribution was therefore Gaussian-like. The error generator used a clock provided by the MUSICAM encoder to synchronize the error pulses with the compressed data stream and to determine the random time interval between single errors. The generator was capable of generating single errors at rates ranging from $5x10^{-5}$ up to 10^{-2} with bursty data clock up to 6.144 MHz or continuous data clock up to 384 kHz. A special module built in the encoder was used for the error insertion.

2.1.4 Generation of FM material

The FM version of the test sequences used in the FM vs MUSICAM comparison was produced by the CBC. A simplified representation of the equipment set-up used is shown in Figure 3. The reference version of the sequences were available on a DAT cassette which was played back on a source DAT recorder. The analog left and right outputs of the source DAT were fed to the input of a stereo coder which generated the FM stereo multiplex. The RF output of the exciter was terminated into a dummy load and also fed to the input of a professional VHF-FM receiver by means of a directional coupler. A variable attenuator was also used to adjust the FM signal level. The VHF-FM receiver was equipped with a conventional 200 kHz IF filter. The demodulated stereo multiplex was then decoded back to analog left and right and recorded on a target DAT. No limiting nor companding of the audio signal was performed leaving full dynamic range to the signal. Band limiting to 15 kHz was performed internally by the stereo coder.

A series of measurements[4] were performed on this FM transmission chain. The frequency response was essentially flat from 50 Hz to 15 kHz. Measurements were done both at 50% and 90% modulation, where +/- 75 kHz deviation is 100% modulation. Total harmonic distortion was better than 0.2% from 50 to 7500 Hz (measured at 90% modulation) and the stereo separation was better than 42.5 dB from 50 Hz to 15 kHz (measured at 100% modulation). Measured FM RMS noise referenced at 400 Hz was -67 dB. These results exceeded the CBC specifications for an operational FM transmission system. Such a high-quality FM signal is in practice rarely available to the general public.

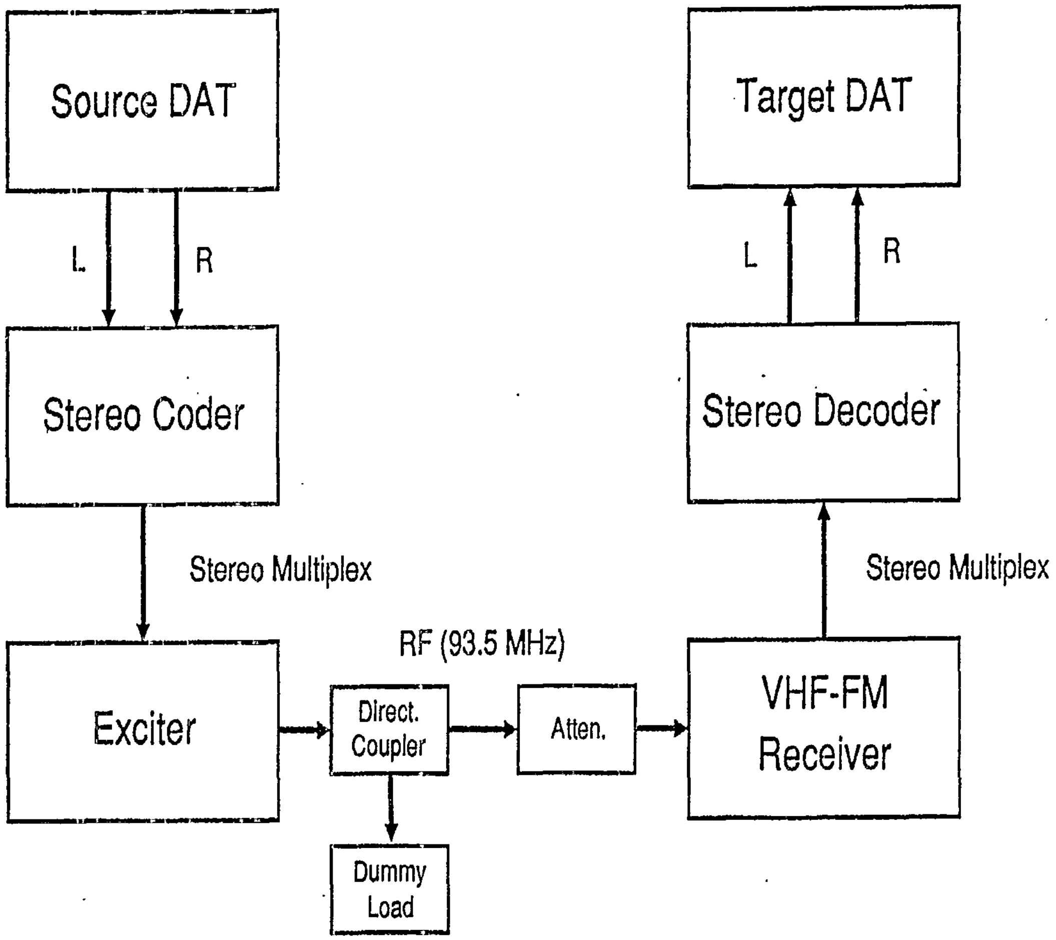


Figure 3 Equipment configuration for generating FM material

2.1.5 Playback system

Figure 4 shows the equipment configuration that was used for the presentation of the test material to the listeners. The headphone tests were done using Stax Lambda Professional headphones with a diffused field equalizer. An expansion block allowed three sets of headphones to be used simultaneously. The loudspeaker tests were done using a pair of JBL 4410 Professional Monitors and the peak sound pressure level was adjusted at 90 dB SPL. For both the headphone and the loudspeaker tests, test sequences were played back on a Sony PCM2500 DAT player.

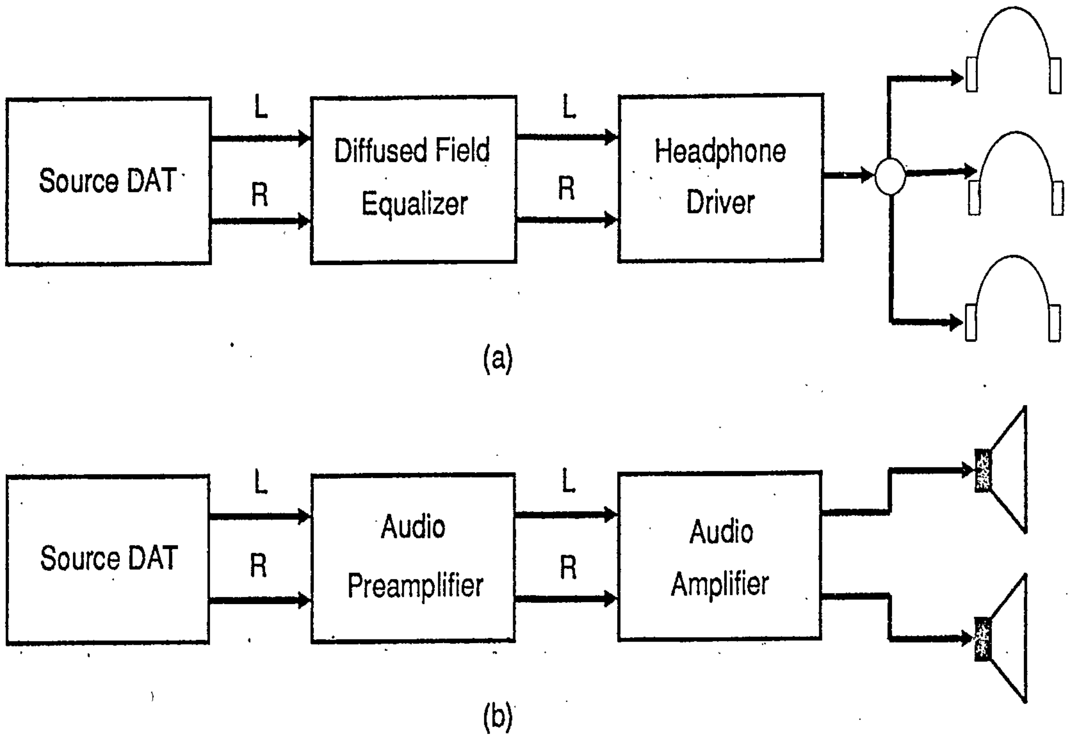


Figure 4 Playback system
a) Headphone Tests
b) Loudspeaker Tests

2.1.6 List of equipment

The following is a list of the various pieces of equipment that were used in the preparation and the presentation of the test material:

MUSICAM MATERIAL

Qty	Item
2	Sony PCM 2500 DAT Recorder (SPDIF, SDIF-2, AES/EBU, analog input/output
1	Panasonic SV-3500 DAT Recorder (SPDIF, analog input/output)
1	Fostex D-20 DAT Recorder (SPDIF, SDIF-2, AES/EBU, analog input/output)
1	Technics SL-P990 CD Player (SPDIF, analog output)
1	Sony DFX2400 Sampling Rate Converter (AES/EBU, SDIF-2)
1	Stax SRM-1/MK-2 Headphone Driver
1	Stax ED-1 Diffused Field Equalizer
3	Stax Lambda Professional Headphones Sets

1	Carver C-2 Audio Preamplifier
1	Carver M-1.01 Audio Amplifier
2	JBL 4410 Professional Monitor Loudspeakers
1	Programmable Error Generator (CRC, custom built)
1	Neumann KU81i Dummy-Head

FM MATERIAL

Qty	Item
2	Sony DTC 1000es DAT Recorder
1	Rohde & Schwarz MSC Standard Stereo Coder
1	Rohde & Schwarz MSDC 2 Standard Stereo Decoder
1	Rohde & Schwarz EU 200 VHF-FM Relay Receiver
1	Harris MS15 Exciter
1	Narda 3020A Bi-directional Coupler
1	Bendix 634N Dummy Load
1	Texscan RA 104 Variable Pad

2.2 Listening room

The loudspeaker tests were conducted in a newly constructed listening room about 7x7 meters in dimension. At the time the bulk of the listening tests were run in November and December of 1990, the room had been acoustically treated with sound absorber panels. The background noise level had a rating of NC-30. Following these tests, modifications to the ventilation system were carried out and the noise level was reduced to NC-25 prior to the second Basic Audio Quality experiment run in April 1991. Further modifications has since been carried out and the room currently meets the NC-20 level between 63 and 4000 Hz and the NC-24 level between 63 and 8000 Hz. Additional work will be carried out to further reduce the background noise level so that the room will achieve somewhere between NC-15 and NC-20.

As will be pointed out later in the report, two experiments were carried out to check the Basic Audio Quality of the MUSICAM system. Both experiments were conducted with both loudspeakers and headphones. No significant differences were found between the results obtained with loudspeakers and with headphones. This is a clear indication that the listening room had no meaningful effect on the outcomes of the experiments reported in this document.

2.3 Listening panel

2.3.1 Composition

About half of the pool of listeners in the six experiments were employees of the CRC. These were mostly scientists, engineers and managers. They came from various, though mostly scientific backgrounds and a number did work in the field of audio. The other half of the listening panel, from outside the CRC, included CBC employees from Montreal and Toronto. These were sound broadcast engineers or professionals working in broadcast studios. A few scientists working in audio research at the National Research Council of Canada (NRC) were also in the listening panel. Another source of listeners was in the general community of Ottawa where volunteers were found from various sources. Many of these were audiophiles or were professionally involved in audio-related work.

While the nature of each experiment was clearly described to each participant, time constraints precluded any attempt at training listeners except for the second of two experiments on Basic Audio Quality assessment described later. For the other five experiments (including the first one on Basic Audio Quality), many listeners participated in more than one experiment. These listeners undoubtedly acquired some degree of increased expertise, both from previous exposures to the actual auditory materials, as well as from familiarization with the experimental procedures and tasks. Even when participants were in more than one experiment, the sequential order of the experiments varied among the listeners, so that any acquired expertise did not accrue to any one experiment more than to any other.

Table 1 shows the number of subjects that participated in each of the identified experiments:

Experiment	Presentation	Number of subjects
Basic Audio Quality #1	Headphones/Loudspeakers	30
Basic Audio Quality #2	Headphones/Loudspeakers	35
Tandem Coding Capability	Loudspeakers	28
Stereo Image Quality	Headphones	33
Monophonic Compatibility	Loudspeakers	19
Robustness to Bit Errors	Loudspeakers	20
FM vs MUSICAM	Headphones	25

Table 1 Number of listeners per experiment

2.3.2 Seashore test

All listeners were administered the Seashore Tests of Musical Talent[5]. This series of test attempts to assess an individual's music related profile in six categories: pitch, loudness, rhythm, time, timbre and tonal memory. It was developed many years ago and has since fallen out of general use, at least in North America. Norms were developed for the series so that any individual can be given a percentile ranking in comparison to the general population on each of the six components. Predictive correlations for the tests, such as success in pursuing musical activities, were at least moderately well-established by the author of the tests, Carl E. Seashore.

Our interest in these measures was as a device to "calibrate" listeners as judges of music-related materials such as we used in many of the experiments reported here. As reading of the results of these experiments will show, the tests did, in some instances, provide categorization that was reflected in the data. These tests do appear to assess perception beyond a matter of sensory acuity. For example, memory for tonal or rhythmic sequences is a capacity that would not be revealed by acuity tests. We are somewhat encouraged that the use of tests of this kind may help in reducing the amount of variance in experimental data. This might facilitate more efficient experimentation and help in getting more reliable answers to experimental questions involving audition of music-related materials. The important conclusions reached in the experiments however would not be modified if the Seashore factor was omitted in the analysis. Alternative tests to the Seashore are described in the literature[6,7].

Strictly speaking, few in the listening panel would qualify as "expert" listeners in the sense of having extensive experience in critical listening to potential digital coding distortions. The majority of these listeners could however be qualified as "skilled" or "experienced" because of their professional background or their special interest in audio. For most of the experiments, we were able to divide the listeners into a high and a low group based on the mean outcome on the six Seashore scales. A mean Seashore tests score of 70 was used as the threshold to define these groups. Most of our "low" Seashore group were still above average in these musical judgement tests in comparison to the general public.

2.4 Data analysis

Each of the experiments were analyzed by standard analysis-of-variance (ANOVA) procedures, and standard sub-tests (e.g. Scheffe and Newman-Keuls comparisons). ANOVA tables are included for the reader familiar with these statistical techniques. However, a detailed understanding of either the analysis or the tables is not necessary. In each of the tables, the reader should only note the definition of the factors as stated at the bottom of each table, the factor column and the probability level column (p-level). In other words, the first and last columns in each table are the essential ones. Each factor considered in isolation of all other factors is called a "main effect" and any of the possible combinations of factors are "interactions". Only those main effects and interactions which achieve a p-level of 0.05 or less can be considered "real" or "significant", i.e. not due to the operation of chance factors. In all cases, the discussion of the experiments clarifies the interpretation falling out of the analyses.

3. Basic Audio Quality

3.1 Purpose

The main purpose of this test was to assess the transparency of the MUSICAM system, that is to examine whether audio materials processed through this system were perceptually different from a reference (unprocessed) version. A secondary purpose was to see whether listeners outcomes on the Seashore Tests of Musical Talents correlated with the ability to make perceptual discrimination of the type under study.

3.2 Test method

Two different experiments were conducted to assess the basic audio quality. They will be referred to as experiment 1 and 2. Both experiments used repeated measures (within-subject) designs where each listener made judgements under all the factors of the experiment in all combinations. As we will see, listeners in the first experiment failed to reliably detect any differences between the reference audio materials and the same material processed through MUSICAM. This outcome differed from the ones shown in the MPEG Audio Test Report[8]. However, there were a number of procedural differences between the MPEG study and our first experiment. To see whether some of these procedural differences were responsible for the differences in outcomes, we conducted the second experiment, changing certain manipulations to make that study more similar to the MPEG one.

In both experiments, the listeners were to judge if they perceived any difference between the reference audio materials and a version processed through the MUSICAM system. Any detected difference was to be considered as an impairment and the listeners were asked to grade the degree of impairment. For this purpose, the double-stimulus (A-B) presentation in conjunction with the five-grade impairment scale recommended in CCIR Rec. 562-2 and described in Table 2 below were used.

5	Imperceptible
4	Perceptible but not annoying
3	Slightly annoying
2	Annoying
1	Very annoying

Table 2 Grading scale for Basic Audio Quality

A trial consisted in the presentation of two stimuli A and B and the listeners were asked to grade B in comparison to A. In experiment 1, stimulus A was always the reference (unprocessed) sequence. Stimulus B was either hidden reference or MUSICAM sequence or a low anchor as shown in Table 3. Low anchors were deliberately impaired sequences obtained by processing the reference sequences 14 times in cascade through the MUSICAM system. On the expectation that impairment in the low anchors would be relatively easy to detect, they provided a check on the sensitivity of the listeners and of the experimental procedure to reveal audible differences. Experiment 2 was identical except that low anchors were not included and only 8 of the 10 audio materials of experiment 1 were used.

Combination	A	B
1	Reference	Hidden Reference
2	Reference	MUSICAM
3	Reference	Low Anchor

Table 3 A-B combinations used in experiment 1

General instructions delivered verbally to the subjects told them about the nature of the MUSICAM coding and provided them a detailed explanation of the use of the grading scale. They were informed that the grading scale was continuous and that they could assign score values with one decimal (e.g. 3.5, 4.7, etc...). A grade of 5.0 was to be given when A and B were perceived as identical. A grade between 4.0 to 4.9 was to be given when a perceptible but not annoying difference was detected in B when compared to A. Grades between 3.0 and 3.9 corresponded to slightly annoying differences and so on (see Table 2 above).

The A-B method used in the present study bears some similarity with the triple-stimuli with hidden reference method (A-B-C) in that both the sequences from the system under test and the hidden references are to be graded. It is different from the A-B-C method in the way the material to be assessed is presented to the listeners. In the A-B method, the hidden reference and the system under test are presented and assessed in two different trials whereas in the A-B-C method they are presented and graded in the same trial. The A-B method was preferred over A-B-C method because of the unreliability of long or medium-term aural memory and because our listening test facilities did not allow the listeners to control switching between the individual sequences. Also, in the way these two methods are used, there is some question about whether the A-B-C method is truly blind to the listener when he/she is informed that B or C is a hidden reference. On the other hand, there is ample assurance that the A-B method is truly blind.

In experiment 1, listening sessions started with a single presentation of the reference version of each of the 10 test sequences described in Table 4 of section 3.3. This was done to allow the subjects to familiarize themselves with the auditory materials. The experiment proper then began and the A-B pairs to be graded were presented. Each of the 10 test sequences was presented in the 3 different A-B combinations described in Table 3 to yield a total of 30 trials. The test sequences were presented in a cyclical order (seq. 1 to 10 in Table 4 below) from trial to trial but the A-B combinations varied in a manner that was unpredictable to the listeners.

In experiment 2, the first eight sequences described in Table 4 were used. Sequences 9 (Fireworks) and 10 (Bass synth.) were excluded because they were found to be totally uninformative in experiment 1. Each listening session in experiment 2 began with a small "training" session which consisted of the presentation of the eight test sequences in A-B pairs. For these eight training trials, the subjects were informed that the first stimulus (A) was the reference and the second one (B) the MUSICAM processed version. The training session was followed by the presentation of the A-B pairs to be graded. Each of the eight audio materials was presented in the two first combinations described in Table 3 (low anchors were not used) and each combination occurred twice in the experiment to yield a total of 32 trials. The test sequences were presented in cyclical order (seq. 1 to 8) from trial to trial but the A-B combinations varied in a manner that was unpredictable to the listeners.

In both experiments, listeners were not informed that the A member of a pair was always an unprocessed reference except, of course, in the training part of experiment 2. A total of 30 listeners took part in the first experiment and 35 in the second one. Both headphones and loudspeakers were used by all listeners in both experiments.

3.3 Test material

The following ten sequences used by the ISO-IEC/MPEG committee for its listening tests were also used here:

Seq. #	Title	Track/Index	Time	Source
1	Suzanne Vega	1	00:22 - 00:42	A&M 395 136-2
2	Tracy Chapman	6	00:36 - 00:57	Elektra 960 774-2
3	Glockenspiel	35/1	00:00 - 00:16	EBU SQAM 422-204-2
4	Ornette Coleman	7	19 s	Dreams 008
5	Castanets	27	00:00 - 00:20	EBU SQAM 422-204-2
6	Male Speech	17/2	54:16 - 54:35	Japan Audio Soc. CD-3
7	Bass Guitar	-	20 s	RR Recording (DAT)
8	Trumpet Concerto	10	05:10 - 05:30	Philips 420 203-2
9	Fireworks	1	00:00 - 00:15	Pierre Verany 788031
10	Bass Synth.	-	25 s	RR Recording (DAT)

Table 4 List of test sequences used for Basic Audio Quality

3.4 Test results

3.4.1 Experiment 1

The results were analyzed using standard analysis-of-variance (ANOVA)[9 to 14] statistical techniques. Detailed comparisons used either or both of Newman-Keuls and Scheffe tests, as indicated in the presentation of results below[9 to 14]. The following table presents the overall outcomes of the ANOVA for experiment 1. (See section 2.4, page 11, for a brief explanation of the important data analysis parameters).

Factors	df Effect	MS Effect	df Error	MS Error	F	p-level
1	1	9.1276	28	9.111960	1.0017	.325466
* 2	1	4.4696	28	1.001494	4.4629	.043686
** 3	2	348.2439	56	2.419252	143.9469	.000000
** 4	9	22.5726	252	.820266	27.5186	.000000
* 1x2	1	4.4693	28	1.0011494	4.4627	.043691
** 1x3	2	16.8689	56	2.419252	6.9728	.001977
2x3	2	.0193	56	.461795	.0418	.959117
1x4	9	.6200	252	.820266	.7559	.657387
** 2x4	9	2.3375	252	.305364	7.6547	.000000
** 3x4	18	13.1743	504	.478251	27.5469	.000000
1x2x3	2	.2755	56	.461795	.5967	.554101
1x2x4	9	.1295	252	.305364	.4240	.921630
√ 1x3x4	18	.7458	504	.478251	1.5594	.066030
** 2x3x4	18	1.9795	504	.270637	7.3140	.000000
1x2x3x4	18	.1347	504	.270737	.4976	.959319

** significant effect at p << .05
* significant effect at p < .05
√ noteworthy effect at p < .07

Factors: 1 = Seashore test (High, Low)
2 = Transducer (Loudspeakers, Headphones)
3 = Coding (Reference, MUSICAM, Low Anchor)
4 = Audio materials (10 items)

Table 5 Basic Audio Quality, Experiment 1, ANOVA Summary

3.4.1.1 Major findings - Transparency

All significant interactions which involved the Seashore factor (namely, 1x3x4, 1x3 and 1x2 in the ANOVA, Table 5) do not impact on the main findings relevant to the transparency of MUSICAM. Discussion of all of these will be postponed until section 3.4.1.2 below.

The most important finding in this experiment concerns the interaction between Coding and Audio materials (3x4 in the ANOVA, Table 5). This is presented in Figure 5. It is apparent there, that there are very few differences to be found between the Reference and MUSICAM grades for any of the Audio materials. On the other hand, the Low anchor level of the Coding factor, appears to receive consistently lower grades on all materials except Fireworks and Bass synth (materials 9 and 10).

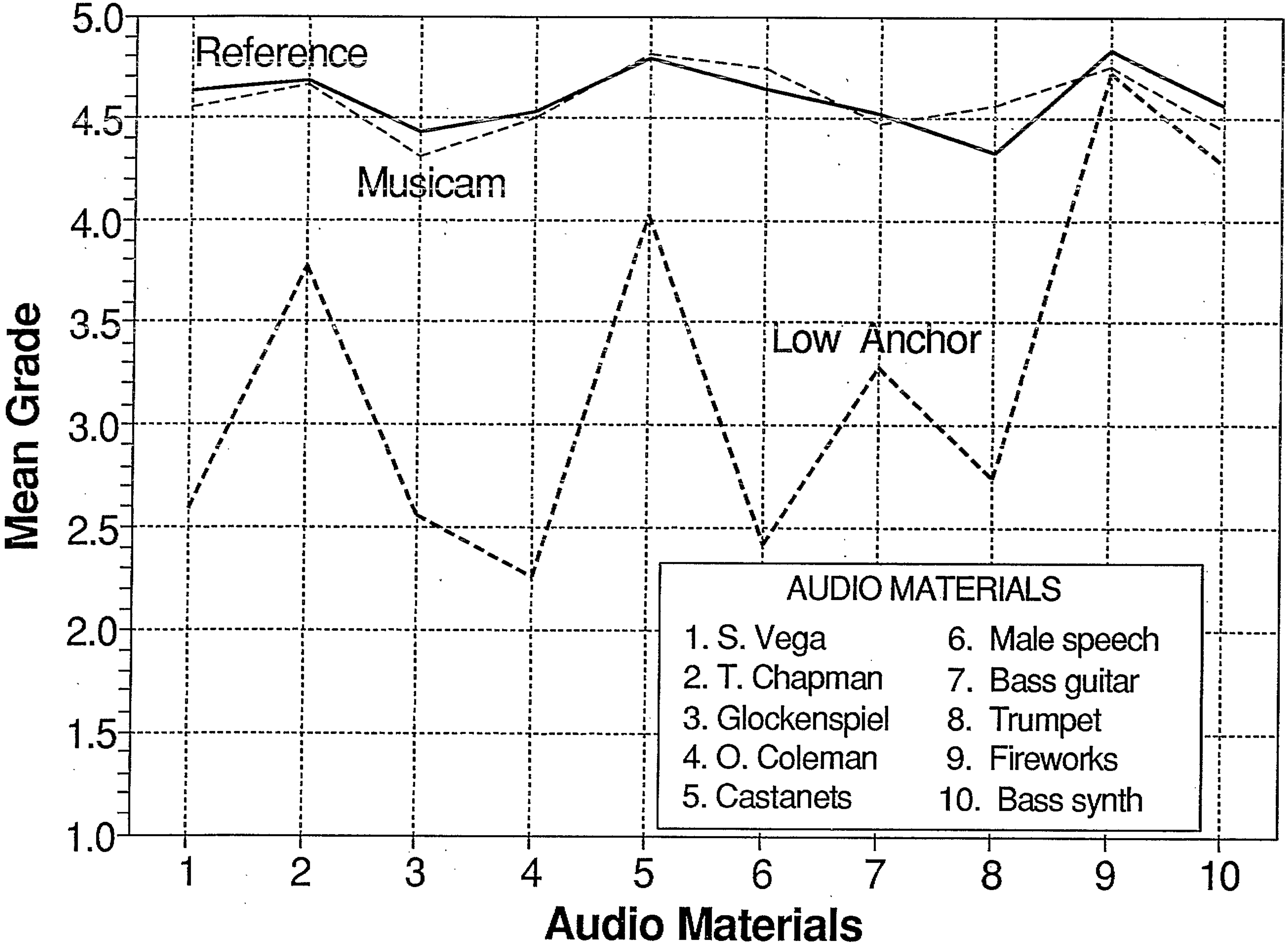
This description is fully confirmed by both Scheffe and Newman-Keuls comparisons. These comparisons show full transparency for MUSICAM coding on each of the Audio materials (no Reference-MUSICAM differences at any Audio material, p < .05). In addition, no differences between the materials within either the Reference or the MUSICAM Coding conditions are significant according to statistical sub-test comparisons not shown in detail here.

For the low anchor samples ("vertical" view), the Scheffe comparisons show that for 7 of the Audio materials, rating differences between these and the comparable Reference and MUSICAM samples are statistically reliable.. The three samples that show no differences from either the Reference or MUSICAM counterparts are the Castanets, Fireworks and Bass synth (materials 5, 9 and 10). The Newman-Keuls tests ascribe a significant difference to the Castanets, but otherwise, the conclusions about Reference and MUSICAM compared to Low anchor are the same.

The patterns for the samples within the Low anchor level ("horizontal" view) are somewhat different for the two comparison procedures. The Newman-Keuls tests place the five materials with the lowest grades in Figure 5 (S. Vega, Glockenspiel, O. Coleman, Male speech and Trumpet) into a single group with no reliable differences among them. The other five materials are statistically different from each other. The Scheffe tests agree with the Newman-Keuls regarding the lowest rated samples and tend to add the Bass guitar to that group. However, the Scheffe tests place the remaining higher ranked samples (T. Chapman, Castanets, Fireworks and Bass synth) in a single undifferentiated group rather than being independent of each other. We will deal at greater length with the implications arising out of the Low anchor findings in our later discussion about critical materials.

Next, we will examine the highly significant three-way interaction involving Transducer, Coding, and Audio materials (2x3x4 in the ANOVA, Table 5). As we will see, this interaction did not prove particularly interesting for the purposes of the experiment. Graphical presentation is not warranted since the interesting aspects of the findings (the 3x4 interaction) were examined above.

Figure 5 - Two-way interaction, Exp.1
Coding by Audio Materials



Detailed comparisons showed that statistical significance of this three-way interaction is entirely due to the differences between the Low anchor outcomes under the two transducers. Newman-Keuls comparisons (p < .05) show that the ratings for four of the 10 Audio materials with Low anchor coding are significantly different between Loudspeaker and Headphone listening. However, two of these materials (Glockenspiel and Trumpet) show reliably HIGHER ratings on headphones compared to loudspeakers, while the other two (Male speech and Bass guitar) show reliably LOWER ratings on Headphones versus Loudspeakers. In other words, the effect obtained, although statistically reliable, lacks generality for two reasons: (1) it involves only 4 of the 10 audio samples, and (2) the direction of the differential effects due to the Transducer factor are in two opposite directions. It is likely, then, that other audio materials not included in this experiment would also show inconsistent directions of differences. Our outcome here does suggest that Headphones are better at revealing distortions on certain Audio materials, while Loudspeakers are better on other Materials. Exploration of this Transducer effect, however, is best done by experiments specifically addressed to systematically manipulating the characteristics of Audio materials in a way that was not part of the present experiment.

Only the Low anchor level of the Coding factor is involved in the 2x3x4 interaction. The main interest in the experiment has to do with Reference versus MUSICAM, neither of which were affected differentially by the Transducer factor. Accordingly, as stated above, this three-way interaction, although statistically reliable and of some interest, proves to be unrelated to the main purposes of the experiment.

The above analyses exhaust most of the interesting outcomes in the experiment. The remaining significant effects shown in the ANOVA (not counting the ones which involve the Seashore factor which are discussed later) support those findings. We will briefly look at these remaining effects without graphic or tabular presentation.

The significant two-way interaction between the Transducer and the Audio materials (interaction 2x4 in the ANOVA, Table 5) was entirely due to two samples, namely Male speech and Bass guitar. For both of these samples, judgments were harsher (lower ratings) under Headphone listening than under Loudspeakers (Newman-Keuls p < .0001 for both samples). But this is simply the algebraic result (across the Coding factor) of the finding already reported in our examination of the three-way interaction of factors 2, 3 and 4. Similarly, the fact that listener ratings under Loudspeakers and under Headphones were virtually identical for each of the other 8 samples provides no additional information over the findings in the same three-way interaction.

The significant main effect for the Audio materials (factor 4 in the ANOVA) completely due to the Low anchor materials. A graphic presentation would simply show the mean algebraic resultant at each sample across the three levels of the Coding factor as seen in Figure 5. This would follow the pattern seen for the Low anchor samples in that figure, with a diminution in the range between peaks and troughs due to the algebraic summation. Detailed statistical comparisons would add no new information about the materials and are not presented.

Equally unsurprising are the significant main effects for both the Coding factor (factor 3) and for the Transducer one (factor 2). Dealing first with factor 3, the Low anchor samples collapsed across all other factors are rated more harshly (mean 3.27) than either the Reference samples (mean 4.60) or the MUSICAM ones (mean 4.58) while the Reference-MUSICAM samples are practically identical. Turning to the main effect of factor 2, the Headphones produced a slightly harsher judgments (mean 4.10) than the Loudspeakers (mean 4.20). This Headphones-Loudspeakers outcome will be clarified below (section 3.4.1.3) when dealing with the Seashore findings.

3.4.1.2 Discussion - Transparency

As we saw above, the major outcome of the experiment is that absolutely no reliable differences were found between ratings for any Reference material and its MUSICAM version. This total lack of difference persists no matter how the data are examined and analyzed. This outcome is at variance with the MPEG study which did report a small but reliable difference indicating that MUSICAM was not fully transparent. In other words, the listeners in the MPEG study, which used the same 10 materials as here, detected minor imperfections. It must be pointed out that the MPEG study used the A-B-C presentation method as opposed to the A-B method used here.

There was at least one additional major difference between the MPEG study and the present one, namely, that we included a multi-pass (14 passes) MUSICAM version for each material. These Low anchor materials might, in fact, have contributed to our failure to find concurrence with the MPEG study. Most of these Low anchor stimuli were found to be discriminable from the Reference and MUSICAM versions. We are suggesting that our listeners might have paid closer attention to tiny differences between the Reference and MUSICAM versions if they had not been distracted by being given the more easily discriminable differences provided by the Low anchor materials. This kind of "series effect", wherein judgments of the magnitude of a difference are strongly affected by the range of differences encountered, has been well-known in behavioral research from the earliest days of psychophysical, and even social-psychological research[15].

To examine this possibility, we performed a second experiment, reported in section 3.4.2, where we used only Reference and MUSICAM versions. In this second experiment, we also provided explicit familiarization, or "training" to try to take into account another difference between the present study and the MPEG one, namely, the fact that the MPEG listeners were generally more experienced in judging coding distortions. By contrast, our listeners were not experienced in this way. Additionally, in experiment 2, we used only 8 of the materials. We dropped two, Fireworks and Bass synth, because in their Low anchor versions they could not be distinguished from their Reference and MUSICAM versions. This was taken as clear indication that they were not good materials for revealing coding distortions, at least for the MUSICAM system.

Apart from the major findings about transparency, we saw some differences among the Audio materials in the absolute grades they received in our data. Such differences among materials were especially evident in the Low anchor condition. These variations are of interest in assessing which among these materials are more and which are less critical for revealing coding distortions. Further discussion from this viewpoint will be elaborated in section 3.4.1.4 below.

3.4.1.3 Major findings - Seashore factor

The main item of interest here is the interaction of the Seashore factor with Coding and Audio materials (1x3x4 in the ANOVA, Table 5). It should be noted that this effect falls just short of the generally accepted probability level of .05. Our conclusions must therefore be offered with some caution. Nonetheless, the outcome is close enough to statistical criteria to warrant attention.

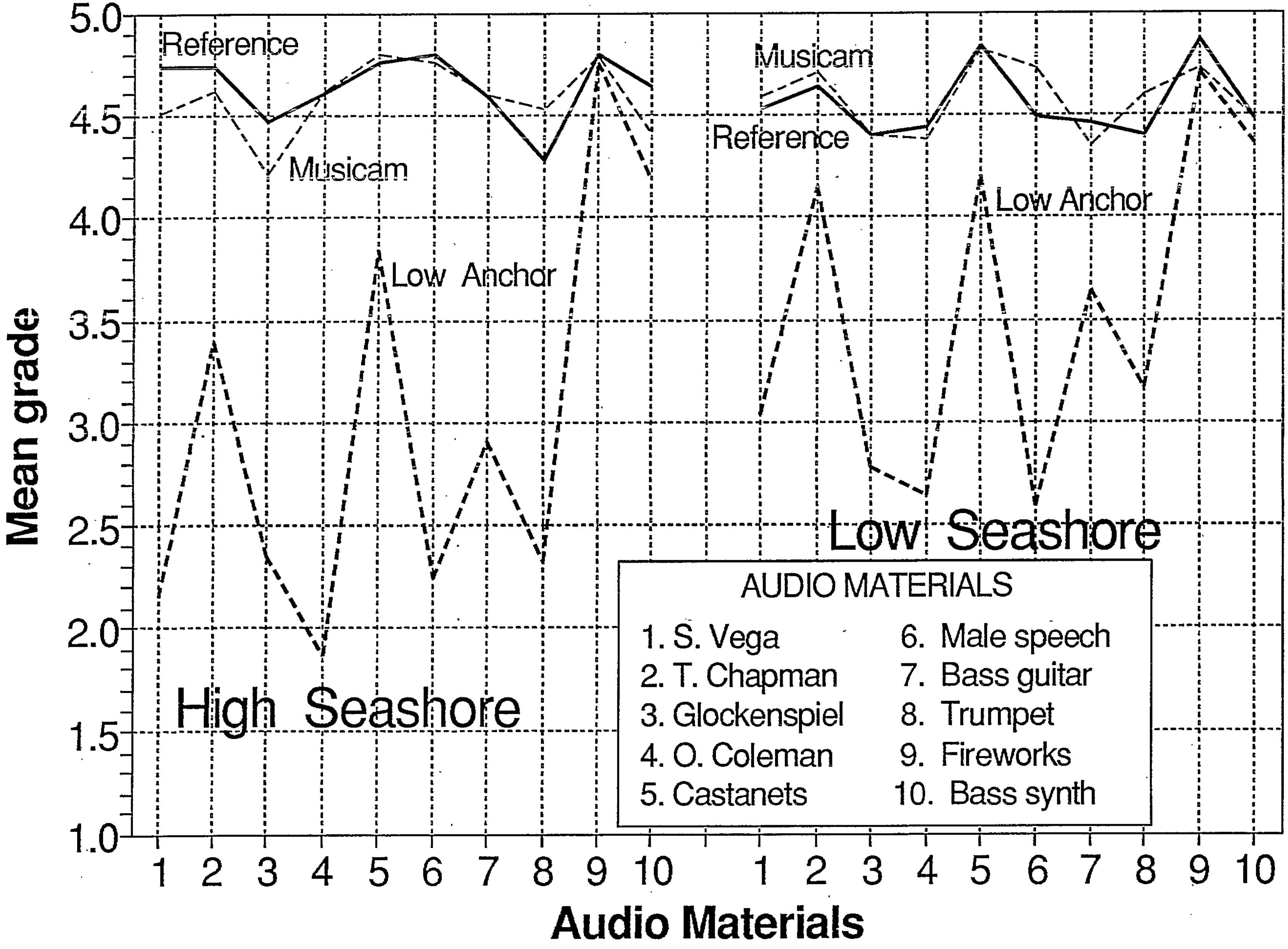
The data for this interaction are shown in Figure 6. This is the same data as was shown in Figure 5 but now differentiated between the Seashore groups. Appropriate sub-analyses (Newman-Keuls) show that interaction is entirely accounted for by the Low anchor materials. Had this not been the case, then our conclusions about the transparency of MUSICAM would need modification. As it is, those firm conclusions remain unchanged in any way by the effects of the Seashore factor.

The figure shows that High Seashore listeners tended to give lower overall ratings to the Low anchor materials than the Low Seashore group. Since the interaction is marginally significant, as discussed above, it is not surprising that Scheffe tests (the most conservative comparison technique) show that these apparent Seashore differences are not significant. The more lenient Newman-Keuls comparisons do show differences. Specifically, 5 of the materials (S. Vega, T. Chapman, O. Coleman, Bass guitar and Trumpet) are reliably rated more harshly by High than by Low Seashore listeners ($p < .05$). No differences were found between the Seashore groups on the remaining five materials (Glockenspiel, Castanets, Male speech, Fireworks and Bass synth).

It appears then (if we go with Newman-Keuls), that the Seashore factor does tend to discriminate between listeners who are harsher judges (High Seashore) and those who are more lenient (Low Seashore) when evaluating severely impaired materials. Since this was evident on only certain of the audio materials, it suggests that those materials may be more critical ones for revealing MUSICAM impairments. This suggest too, that if the experimental materials had exclusively included only more highly critical sequences, then the Seashore factor might have emerged more strongly as a significant factor in the experiment. Further discussion of the implications of this experiment for critical materials will be presented in section 3.4.1.4.

Since High Seashore listeners were harsher judges of certain of the Low anchor materials, we might argue that they are more sensitive listeners than the Low Seashore group. If this is true, then we may speculate that they should have been able to uncover Reference-MUSICAM differences if they were there to be detected. The fact that they did not do so perhaps strengthens the conclusion that MUSICAM is totally transparent to listeners such as those in our experiment, even to the ones that are high in sensitivity to musical materials. In any case, both because of

Figure 6 - Three-way Interaction, Exp.1
Seashore by Coding by Audio Materials



the detectability of distortions in the Low anchor condition by all our listeners, as well as because of the differential between the Seashore groups, our experiment cannot be faulted for lack of sensitivity to revealing coding distortions.

Turning to the highly significant interaction between the Seashore factor and Coding (1x3 in the ANOVA), this is the identical data to that in Figure 6 but collapsed across the Audio materials factor. As would be expected, this simply shows, once again, that there were differences between the two Seashore groups, all due exclusively to the Low anchor materials where the High Seashore listeners assigned harsher grades than did the Low Seashore group. Where the three-way interaction which included the Audio materials factor (1x3x4, discussed above) was statistically weak because only 5 of the materials showed a differential effect due to the Seashore factor, here, collapsed across the ten materials, the interaction emerges as quite reliable on statistical grounds.

The final significant effect involving the Seashore factor is the one which also includes the Transducer factor (1x2 in the ANOVA). This does provide a new finding among our analyses. We saw previously (at the end of section 3.4.1.1) that overall (main effect of factor 2 in the ANOVA), the Headphones yielded slightly harsher judgments (mean 4.10) than did the Loudspeakers (mean 4.20). It now turns out that when the Seashore factor is taken into account, this effect is entirely due to the Low Seashore group who showed lenient judgments on Loudspeakers (mean 4.32, $p < .05$ in comparison to all the other three data points). On Headphones, this group was identical statistically (mean 4.12) to the High Seashore group on both Loudspeakers and on Headphones (both means = 4.08). And so, independent of all other factors, the High Seashore group tended to give the same judgments regardless of Transducer, while the Low Seashore group was reliably more lenient on Loudspeakers.

If we may agree that headphones provide a situation where finer details of musical passages can be heard more clearly because of the absence of room effects which tend to blur detail, then it would follow that the High Seashore group appeared to be able to discount such blurring in arriving at judgments of the materials while the Low Seashore group were swayed by the room effects. If this explanation is true, then it suggests that better listeners can detect distortions even under sub-optimal listening conditions, while less discriminating listeners need better conditions in order to detect these distortions. However, this may be overinterpreting our finding, since none of the three-way interactions involving both the Seashore factor and the Transducer one were significant (i.e., neither 1x2x3 nor 1x2x4 in the ANOVA even approached significance).

Summarizing the Seashore factor findings, we saw that this factor tented to differentiate listeners who were harsher judges of severely distorted (Low anchor) materials (High Seashore, Figure 6). The same High Seashore listeners were more consistent in their judgements of quality regardless of Transducer (Headphones versus Loudspeakers). These findings are interesting but the major conclusions of the experiment would be unaltered if the Seashore factor was not included in the analysis.

3.4.1.4 Notes on critical materials

Throughout our experiment, no differences were detected between any Reference material and its MUSICAM coded version. Those results alone make it impossible to make any statements about critical materials. However, we did get interesting differences among materials in the Low anchor coding (see Figures 5 and 6). The most striking effect was the failure to find differences between either the Reference or the MUSICAM version on the one hand, and the Low anchor version on the other hand for two materials, namely Fireworks and Bass synth. Since these two materials were not perceptually degraded even with 14-pass MUSICAM, we can safely discard them as not critical for revealing coding distortions at least for the MUSICAM system. For this reason, we did not use these materials in the 2nd experiment on basic audio quality (section 3.4.2).

As for the remaining eight materials, two of them, namely T. Chapman and Castanets, while both significantly different from their Reference/MUSICAM versions, received high ratings compared to the other six materials within the Low anchor versions and were significantly different from those six. Accordingly, T. Chapman and Castanets are both candidates for rejection as critical materials for MUSICAM. The Bass guitar, also received relatively high ratings although lower than T. Chapman and Castanets, and it was significantly different from the lowest five, at least in some comparisons. So it might be regarded as marginal in its ranking along the critical continuum. The remaining five emerge as the most critical materials - S. Vega, Glockenspiel, O. Coleman, Male speech, and Trumpet.

We will not speculate here on why some materials emerged as more critical than others. Such speculation would be fruitful only with further experimental evidence. The Low anchor manipulation, however, emerges as a tool for investigation in exploring questions of this type.

3.4.2 Experiment 2

The results were analyzed using standard analysis-of-variance (ANOVA) statistical techniques. Table 6 presents the overall outcomes of the ANOVA for experiment 2. (See section 2.4, page 11, for a brief explanation of the important data analysis parameters).

3.4.2.1 Major findings - Transparency

The most important outcomes for the major purpose of this experiment is shown by the complete lack of significance for the Coding factor (main effect of factor 3 in the ANOVA, Table 6) and the almost complete absence of significant interactions involving coding. Despite the lack of distracting Low anchor materials such as were present in experiment 1, and despite the initial "training", listeners were completely unable to judge MUSICAM materials as different from the Reference one. We must examine the significant interactions before this conclusion is fully substantiated. But the reader may anticipate that the conclusion about the transparency of MUSICAM is not modified in any way by those interactions.

Factor	Effect	MS Effect	df Error	MS Error	F	p-level
1	1	1.214645	33	6.161872	.197123	.659951
2	1	.041065	33	.440490	.093225	.762034
3	1	.001552	33	.071978	.021569	.884134
** 4	7	3.860030	231	.554563	6.960494	.000000
1x2	1	.129413	33	.440490	.293793	.591442
* 1x3	1	.353767	33	.071978	4.914900	.033628
2x3	1	.443219	33	.251373	1.763195	.193339
1x4	7	.386372	231	.554563	.696714	.674824
2x4	7	.109916	231	.165943	.662373	.703784
3x4	7	.205833	231	.116423	1.767980	.094630
1x2x3	1	.451117	33	.251373	1.794612	.189518
1x2x4	7	.183396	231	.165943	1.105178	.360592
1x3x4	7	.174345	231	.116423	1.497516	.168975
√ 2x3x4	7	.173688	231	.088595	1.960468	.061363
1x2x3x4	7	.085354	231	.088595	.963412	.458728

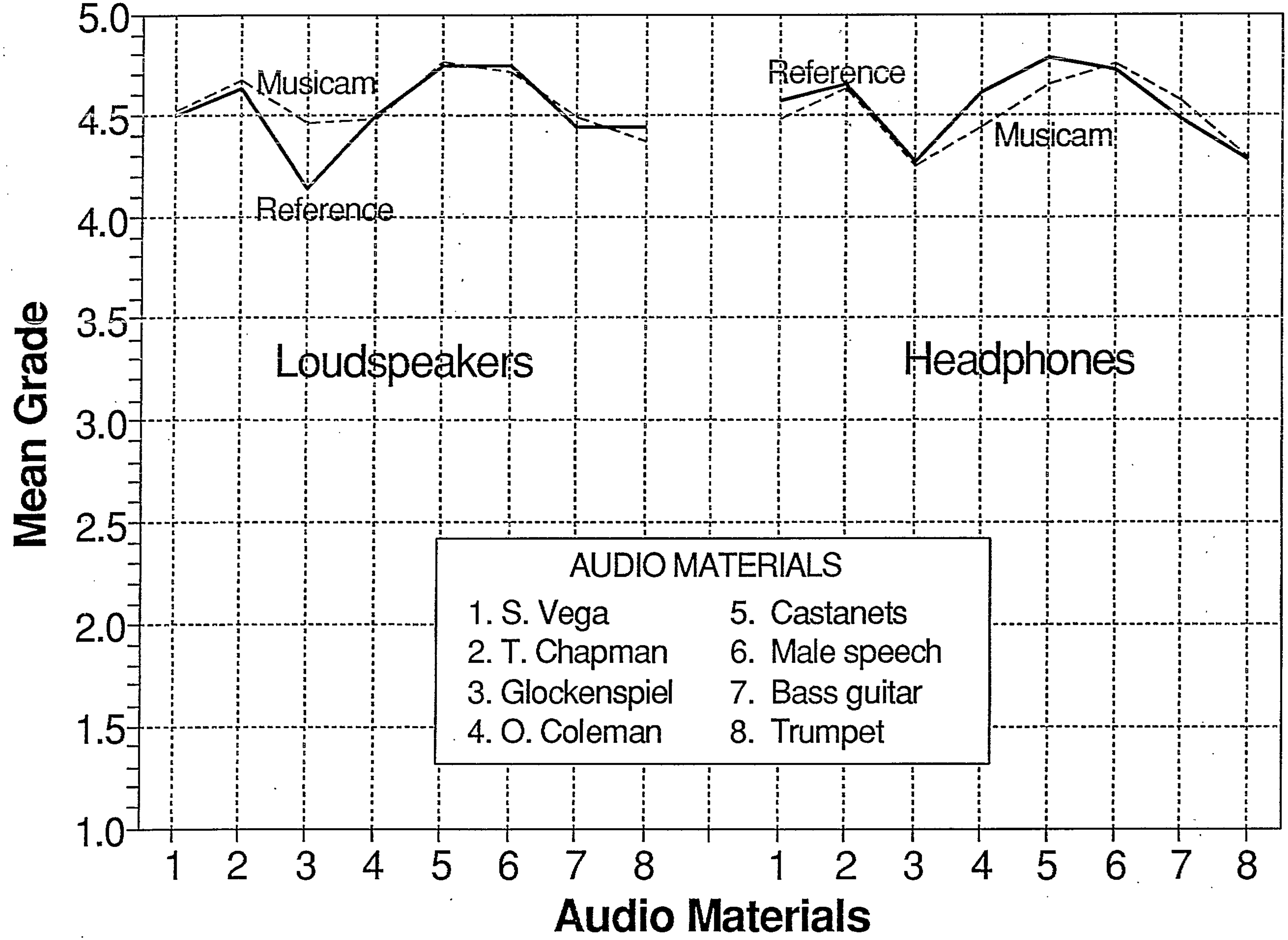
** significant effect at p << .05
* significant effect at p < .05
√ noteworthy effect at p < .07

<u>Factors:</u> 1 = Seashore test (High, Low)
2 = Transducer (Loudspeakers, Headphones)
3 = Coding (Reference, MUSICAM)
4 = Audio materials (8 items)

Table 6 Basic Audio Quality, Experiment 2, ANOVA Summary

While the 2x3x4 interaction falls short, at p < .07, of the generally accepted level of p < .05, it is close enough to be of interest and is presented in Figure 7 below. As might be expected, the conservative Scheffe comparison tests reveal no significant differences. The Newman-Keuls tests, however, single out the largest difference apparent in the figure - that between the Reference and MUSICAM versions for the Glockenspiel on Loudspeaker listening - as significant at p < .001. This is the first and only time, in either the first experiment on basic audio quality, or in this one, that any difference was found between Reference and MUSICAM coding. It should be noted, of course, that the direction of the difference shows that the MUSICAM version of this Audio material received a higher rating than the Reference one; however, there can be no particular significance attached to this "reversal" as such in light of the tenuous nature of this on statistical grounds. We reiterate that the difference we note here is merely suggested but is not confirmed statistically because the interaction does not meet the accepted level of significance.

Figure 7 - Three-way Interaction, Exp.2
Transducer by Coding by Audio Materials



3.4.2.2 Major findings - Seashore factor

The interaction between the Seashore and Coding factors (1x3 in the ANOVA, Table 6) is significant at p < .05. Again, as in experiment 1, the Seashore factor emerges as a reliable differentiator among listener ratings. Subsidiary analysis (Scheffe) shows that effect is entirely due to a higher score for the Reference codings between High and Low Seashore listeners. The effect is tiny (a difference of 0.1 in the score) and so has little real meaning. We merely note its presence. It bears no relationship to the transparency issue.

3.4.2.3 Other findings - Audio materials

The Audio materials effect (main effect of factor 4, ANOVA Table 6) simply collapses the data shown in Figure 7 across both the Transducer and the Coding factors and so graphical presentation would be largely redundant. The Glockenspiel and Trumpet materials got the lowest ratings; T. Chapman, Castanets and Male speech got the highest ones; and S. Vega, O. Coleman and Bass guitar were in the middle. As one would expect, the statistical significance of the effect is due to the differences between the three highest samples (means of 4.65, 4.73 and 4.73) and the two lowest ones (means of 4.28 and 4.35). The rating difference between the lowest of the high materials and the highest of the low ones is 0.3.

3.4.2.4 Discussion

Despite the omission of the Low anchor manipulation in this experiment, and despite the training session where the Coding nature of the samples was clearly identified, listeners still were unable to detect any significant differences between Reference and MUSICAM samples. This was true except for the Glockenspiel sample on Loudspeaker listening. But this one difference only approached marginal significance and does not change the basic generality about the apparent transparency of MUSICAM shown in the experiment.

4. Stereophonic Image Quality

4.1 Purpose

The purpose of this test was to evaluate the ability of the MUSICAM system to provide a stereophonic image subjectively identical to that of a reference stereo programme.

4.2 Test method

This test was done with headphones and a series of reference dummy-head recordings (Neumann KU81i) produced at the CRC specifically for this experiment. These reference sequences consisted in a series of percussive sounds emitted at the following 9 discrete positions in front of the dummy-head: -90°,-67.5°,-45°,-22.5°,0°,22.5°,45°,67.5° and 90°. As shown in Figure 8, 0° was in front of the dummy-head, the negative angles to its left and the positive angles to its right. These reference sequences were coded and decoded with the MUSICAM system.

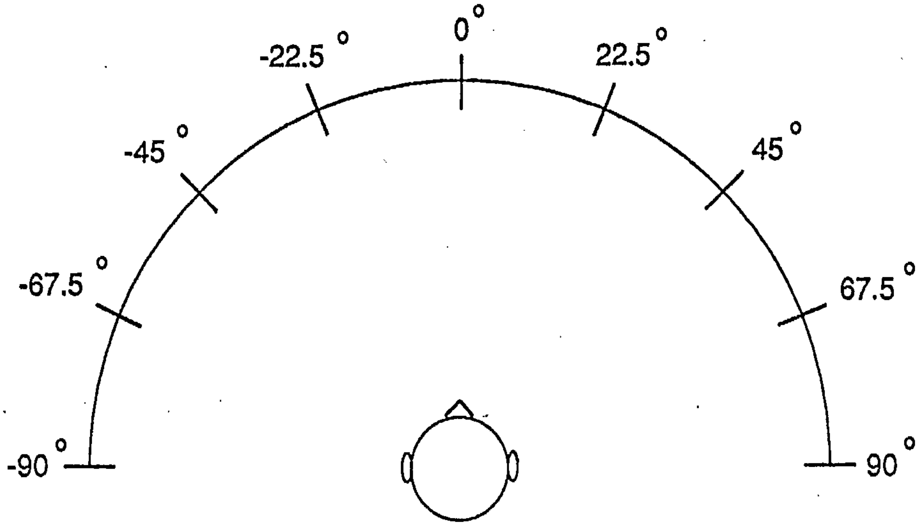


Figure 8 Recording positions of reference stimuli

The reference and the MUSICAM sequences were presented individually (in separate trials) to the 33 listeners in an order that was unpredictable to the listeners. Each reference and MUSICAM sequence was presented 4 times (in 4 separate trials) to yield a total of 72 trials (4x9 reference stimuli + 4x9 MUSICAM stimuli). At each trial, the listener was asked to identify the perceived location of the phantom image with the help of the diagram shown in Figure 9.

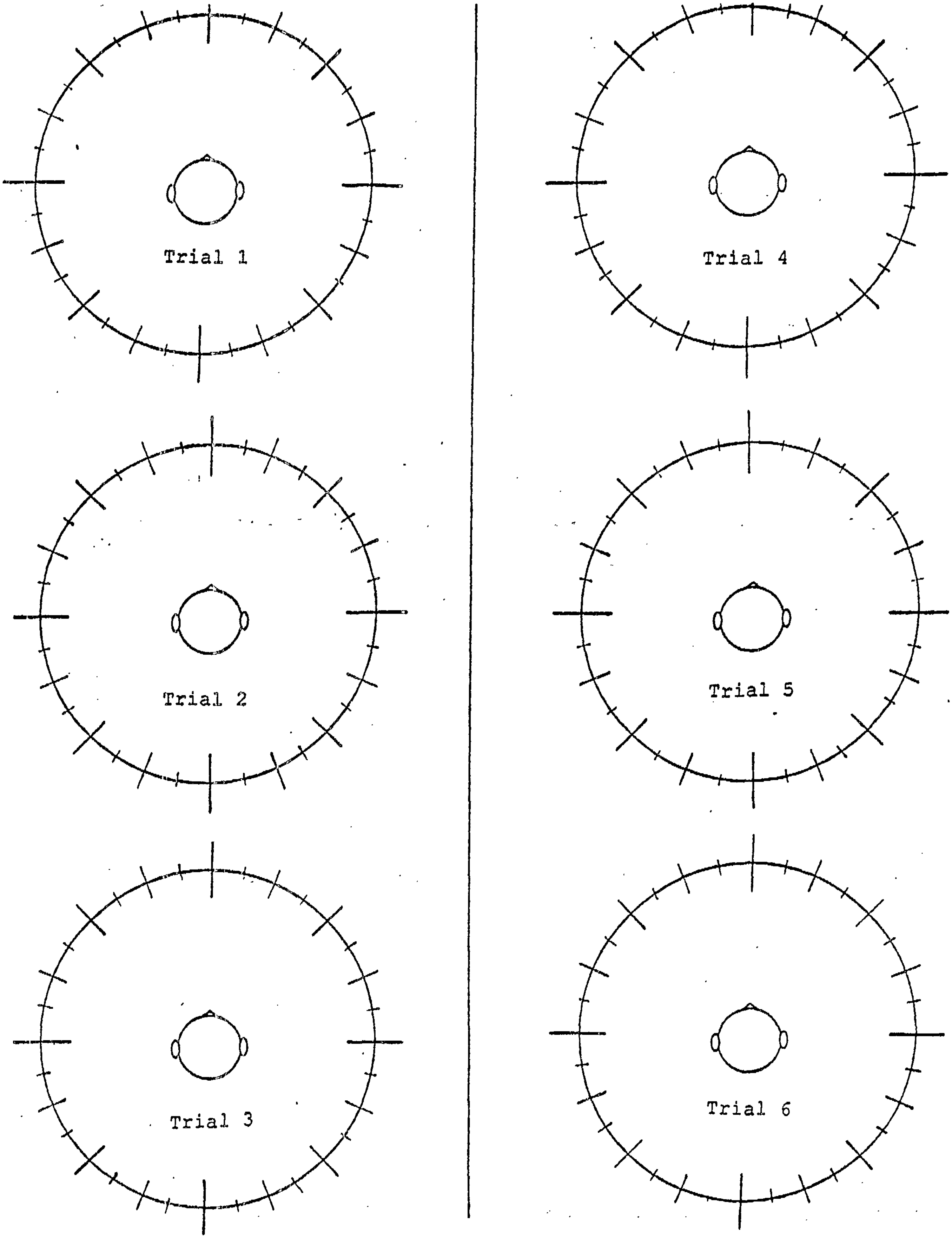


Figure 9 Sample of scoring sheet for the Stereophonic Image experiment

4.3 Test material

A series of 9 percussive sound reference sequences emitted at 9 discrete positions in front of a Neumann KU81i dummy-head. The positions where the sound source was located are shown in Figure 8.

4.4 Data transformation

It is well-known that mirror-image distortions can often occur in the perception of dummy head recordings. Thus, an event at an objective location straight ahead of the dummy-head can be heard as one that was straight behind, and one that was located, say, at 80 degrees to the left of straight ahead, can be heard as 100 degrees left. Inspection of our data showed that this phenomena was strongly in operation in our experiment. Pilot tests done before the experiment was run let us anticipate that this would occur. And so, one of the reasons for using four presentations of the identical recorded location for each of the Reference and MUSICAM versions for each listener was so that the mirroring effect would be properly captured.

Since all the objective locations in the recordings were at the front of the dummy-head, from -90 degrees to +90 degrees left to right from straight ahead (0 degrees), then each of the listeners' observations that fell to the rear of the -90 to +90 line were transformed into their mirror location; a perceived location 123.75 degrees to the right of straight ahead was treated as 56.25 degrees right; one at 135 degrees to the left of straight ahead became 45 degrees left, and so on. The data on the locations that were perceived to the front of the -90 to +90 line were not altered.

Mirroring phenomenon were not the object of study in this experiment, of course; rather the transparency of MUSICAM coding in binaural localization was the issue. While it is obvious that our mirror transformation would reduce the variance in the data and would also uncover the underlying degree of accuracy in localization of binaurally recorded material, we will first examine the frequency of mirroring between Reference and MUSICAM materials.

We counted the number of observations that needed mirror transformation at each of the nine locations for each of the four Reference and MUSICAM encoded events for each listener and converted these to a percentage. The range, then, was from 0% (no mirroring) to 100% (all 4 observations showing mirroring). Then we subjected these data to an ANOVA, with coding (Reference versus MUSICAM) as one factor and objective location (9 positions) as a second one. A significant two-way interaction between these factors would clearly show that the number of scores that required mirror transformation was different for Reference than for MUSICAM material at the different spatial locations. In fact, this interaction was far from significant ($p > .35$). Furthermore, there was no main effect of coding ($p > .40$). Together, these outcomes are strong evidence that considering the amount of mirroring both at each objective location and across all the locations, the MUSICAM encoded materials were not at all different from the Reference materials.

For general interest we can report that the main effect of objective location in this analysis was highly significant (p < .001). This showed that amount of mirroring varied reliably across locations (totally independent, as we saw above, of coding). The highest amount of mirroring was at the +23 degree location (some 67% of the observations); next was the straight ahead (0 degrees) location (64%). The decline was monotonic on both sides of +23 degrees (except for a minor jump at -90 degrees) to lows of 38% and 41% at the lowest points at the right and left.

A second question about mirroring frequency is whether the number of scores that required mirroring treatment was different between Reference and MUSICAM for different individuals. We examined this by calculating the mean across spatial locations for the Reference and for the MUSICAM materials for each individual listener, using the same data as we subjected to the ANOVA described above. A very high correlation was found between these Reference versus MUSICAM scores (Pearson r = 0.93). This clearly shows that the number of scores that required mirror transformations was almost identical for the two codings within individual listener data. The means for the Reference and MUSICAM scores across individuals are 52.4% and 51.3%; the standard deviations are 27.0% and 26.8%.

We believe that both the ANOVA and correlation data presented here establish clearly that, insofar as the number of data points that were subjected to mirror transformations is concerned, Reference and MUSICAM materials were identical. This is our first evidence favouring a conclusion that MUSICAM encoding on binaural materials is transparent since the amount and distribution of mirroring was virtually the same for the MUSICAM as for the Reference materials.

4.5 Test results

The transformed data for the stereophonic image quality test were analyzed using standard analysis-of-variance (ANOVA) techniques. Table 7 presents the overall outcomes of this ANOVA. (See section 2.4, page 11, for a brief explanation of the important data analysis parameters).

The ANOVA clearly shows that none of the interactions had significant effects. Among other things, this means that the Seashore factor was totally irrelevant in this experiment. This is not surprising, since there is no reason to believe that competence in handling musical materials is related to localization in auditory space. More important, there was no relationship at all between Spatial location and the Coding factor (2x3 interaction) so that the two Coding variables did not have different effects as a function of objective location.

However, there were two significant main effects, one for Spatial location (factor 3) and the second one for Coding (factor 2). At first blush the latter would appear to indicate that there was an overall difference in the spatial locations for Reference versus MUSICAM materials, a clear signal that MUSICAM was NOT transparent. However, an examination of the two means that comprise the main effect shows that such a conclusion is not at all justified. The overall mean

Factors	df Effect	MS Effect	df Error	MS Error	F	p-level
1	1	51.2	31	218.2155	.235	.631394
* 2	1	278.0	31	60.2676	4.613	.039660
** 3	8	285968.0	248	269.6446	1060.538	.000000
1x2	1	2.8	31	60.2676	.047	.829460
1x3	8	208.6	248	269.6446	.774	.626298
2x3	8	43.3	248	71.2080	.608	.771226
1x2x3	8	96.8	248	71.2080	1.360	.214780

** significant effect at p << .05
* significant effect at p < .05

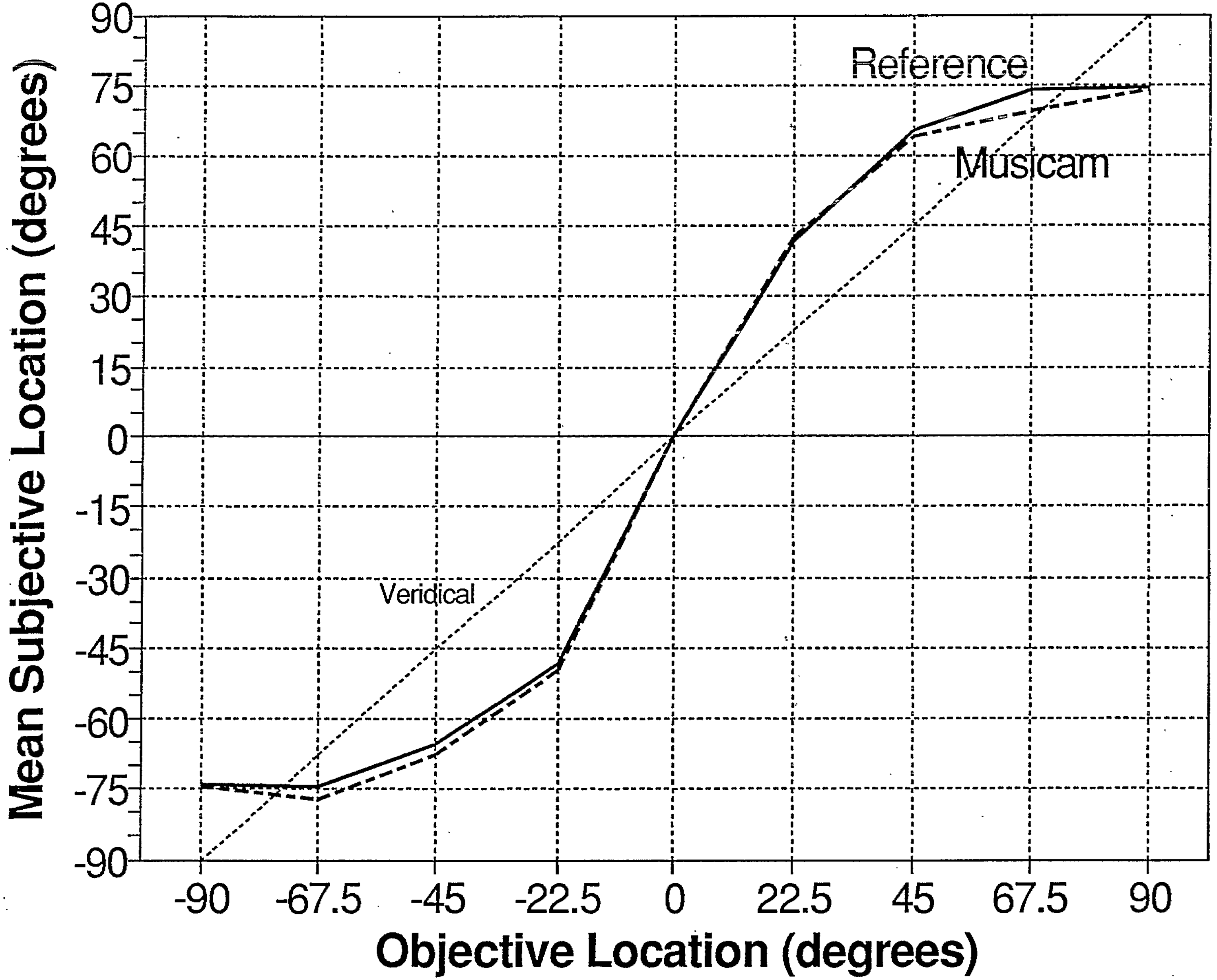
Factors: 1 = Seashore test (High, Low)
2 = Coding (Reference, MUSICAM)
3 = Spatial location (9 positions)

Table 7 Stereophonic Image Quality, Localization means, ANOVA Summary

for the Reference locations was -0.73 degrees, a shade less than 1 degree to the left of straight ahead. The significantly different MUSICAM mean was -2.10 degrees, a trivial 1.37 degrees further left than the Reference mean. The finest resolution our data could produce was 11.25 degrees (this was what the listeners were allowed to discriminate within objective differences of 22.5 degrees). This is an indication that the variances, and hence the error terms, in our data were extremely tiny so that this obvious "noise" appeared as a "signal". This means that our experiment was extremely efficient, and that fewer than the 33 listeners used were needed to reveal the full effects of our factors.

The main effect of Spatial location is shown in Figure 10. Both the Reference and MUSICAM data are shown there and it is obvious that these two codings yielded almost identical outcomes (hence no 2x3 interaction in the ANOVA). For comparison, a "veridical" line which plots what perfect correspondence between subjective and objective locations would look like, is also shown. It is evident that, generally, there is an "overshooting" tendency in subjective locations, both to the left and to the right of 0 degrees, and that this tendency lessens as 90 degrees is approached. Newman-Keuls comparisons (p < .05) show that 7 of the 9 locations are reliably discriminated from each other, the two adjacent ones at the extreme left being statistically identical to each other, as are the two adjacent ones at the extreme right. Scheffe comparisons discriminate 5 points rather than 7 as different, adding the immediately neighbouring points to each of the non-different ones at both the left and right ends.

Figure 10 - Subjective vs Objective Locations



It should be noted that the judged MUSICAM locations tend to overshoot slightly MORE than do Reference ones to the left of 0 degrees, while to the right, judgments under MUSICAM overshoot slightly LESS than the Reference ones. It is the net algebraic effect of these tiny tendencies that yielded the significant, but erroneous main effect of coding (factor 2 in the ANOVA) which we discussed previously.

We also analyzed the data without the mirror transformations. Just as the data presented here, that analysis showed complete transparency for the coding factor too, with virtual identity between MUSICAM an Reference across the locations. Without the mirror corrections, the variances were, not surprisingly, extremely large. Thus, only 3 significantly different locations were revealed, with all 4 points to the left of 0 degrees, and the four points to the right being identical. Both of these groups were different from the straight ahead position. A graph of this untransformed data would conceal the reasonable localization accuracy that we see in Figure 10, since without mirror corrections, the judgments tended to show all 4 points to the left of 0 degrees at about -100 degrees, and all the ones to the right at about +100 degrees. The 0 degree location, seen here as highly veridical, showed up as a mean of about +70 degrees. In other words, the localization accuracy is completely obscured by mirroring if that strong tendency is not taken into account in the treatment of the data.

Our very firm conclusion, even if the analysis on untransformed data is used, is that MUSICAM is fully transparent in its treatment of binaural localization cues recorded on the dummy-head used here.

5. Monophonic compatibility

5.1 Purpose

The purpose of this test was to evaluate the ability of the MUSICAM system to provide a monophonic reproduction (resulting from simple mixing of left and right independently coded signals) of a stereo programme that is subjectively identical to the monophonic reproduction of a Reference stereo programme. In other words, besides the intrinsic loss of the stereo image caused by the mixing process and a possible small impairment caused by the MUSICAM coding process, the mixing operation should not introduce additional degradation.

5.2 Test method

The double-stimuli (A-B) presentation with five-grade impairment scale method described in section 3.2 was used. Each of the test sequences was presented to the listeners in the 4 different A-B combinations described in Table 8 below.

Combination	A	B	Mode
1	Reference	Hidden Reference	Stereo
2	Reference	MUSICAM	Stereo
3	Reference	Hidden Reference	Mono
4	Reference	MUSICAM	Mono

Table 8 A-B combinations used for Monophonic Compatibility

Listeners were thus asked to grade both hidden references and MUSICAM material presented both in stereo and mono. If any differences obtained between combinations 3 and 4 (Table 8) are greater than differences between combinations 1 and 2, then MUSICAM would be said to suffer impairment in the monophonic mode. If any such differences are equivalent, then MUSICAM would be assumed to be compatible with monophonic reproduction.

Each of the five test sequences described in section 5.3 was presented to the listeners in the four different combinations described in Table 8 above to yield a total of 20 trials. The five test sequences were presented in cyclical order (seq. 1 to 5) from trial to trial but the A-B combinations (Table 9) varied in a way that was unpredictable to the listeners. This experiment was presented over loudspeakers and used 19 listeners.

5.3 Test material

The following five test sequences were used:

Seq. #	Title	Track/Index	Time	Source
1	Soprano	61/1	33:00 - 51:00	EBU SQAM 422-204-2
2	Brass ensemble	66/1	00:00 - 00:14	EBU SQAM 422-204-2
3	Wind ensemble	67/1	00:00 - 00:14	EBU SQAM 422-204-2
4	ABBA	69/1	00:00 - 00:15	EBU SQAM 422-204-2
5	Cello & violin duo	5/1	00:00 - 00:15	Japan Audio Soc. CD-6

Table 9 List of test sequences for Monophonic Compatibility

5.4 Test Results

The following table present the ANOVA for the outcome of this experiment. (See section 2.4, page 11, for a brief explanation of the important data analysis parameters).

Factors	df Effect	MS Effect	df Error	MS Error	F	p-level
* 1	4	1.012132	72	.326714	3.097914	.020681
2	1	.063187	18	.224184	.281855	.601978
3	1	.085506	18	.092056	.928853	.347931
1x2	4	.217335	72	.180760	1.202346	.317314
1x3	4	.067414	72	.113061	.596266	.666482
2x3	1	.247607	18	.398605	.621183	.440865
1x2x3	4	.118596	72	.120773	.981980	.422902

* significant effect at p < .05

Factors: 1 = Audio materials (5 items)
2 = Coding (Reference, MUSICAM)
3 = Mode (Mono, Stereo)

Table 10 Monophonic Compatibility, ANOVA Summary

As is clear in the ANOVA table, the only factor which emerged as significant in this experiment was the overall rating of the Audio materials (main effect of factor 1) independent of all other factors. This was all due to a difference between only two of the five samples, namely the Brass Ensemble and the Wind Ensemble. The latter receiving a lower overall rating, entirely independent of any other conditions, of 4.84; the Brass Ensemble received 4.53. This fact is of no consequence for the experimental question and will not be dealt with any further.

The major analysis item is the interaction between factor 2 (Coding) and factor 3 (Mode). If this were significant, it would indicate that a different relationship between the grades for Reference and MUSICAM was found in Stereo mode than in Mono mode. Since this interaction is far from significant ($p > 0.40$), we may conclude that MUSICAM under monophonic reproduction fares just as well as Reference materials in this mode. In other words, MUSICAM appears to be fully compatible in monophonic reproduction.

The only caution to this conclusion is that, apart from the inconsequential difference within factor 1 (Audio materials), the experiment produced no differences whatever. And so, it may be argued that the failure to find a Coding by Mode interaction (factor 2 by 3) may have been due to insensitivity of the experiment for revealing the type of differences under investigation. For instance, the test sequences used might not have been critical for the purposes of the experiment.

We temper our conclusion accordingly. <u>Although the experiment supports monophonic compatibility for MUSICAM, additional experimentation is needed before a stronger conclusion can be made.</u>

6. Robustness to bit errors

6.1 Purpose

The purpose of this test was to determine the failure characteristic of the MUSICAM system in the presence of random bit errors. The failure characteristic is the way the audio quality deteriorates in the presence of increasing bit errors.

6.2 Test method

The double-stimulus (A-B) presentation method with the five-grade impairment scale described in section 3.2 was used. Each test sequence was presented to the listeners in the 7 different A-B combinations described in Table 11 below:

Combination	A	B
1	Reference	Hidden Reference
2	Reference	MUSICAM No error
3	Reference	MUSICAM + BER of 5x10^{-5}
4	Reference	MUSICAM + BER of 1x10^{-4}
5	Reference	MUSICAM + BER of 5x10^{-4}
6	Reference	MUSICAM + BER of 1x10^{-3}
7	Reference	MUSICAM + BER of 5x10^{-3}

Table 11 A-B combinations used for Robustness to Bit Errors

The MUSICAM material corrupted with bit errors was generated with the method described in section 2.1.3. Random single errors were generated at the bit-rates shown in Table 11 above. Each of the 3 test sequences described in section 6.3 was presented to the listeners in the 7 combinations shown in Table 11 to yield a total of 21 trials. The test sequences were presented in cyclical order (seq. 1 to 3) from trial to trial but the A-B combinations varied in a manner unpredictable to the listeners. Loudspeakers were the only transducer used and 20 listeners took part in the experiment.

6.3 Test material

The following subset of the ISO-IEC/MPEG test sequences were used:

Seq. #	Title	Track/Index	Time	Source
1	Suzanne Vega	1	00:22 - 00:42	A&M 395 136-2
2	Glockenspiel	35/1	00:00 - 00:16	EBU SQAM 422-204-2
3	Male Speech	17/2	54:16 - 54:35	Japan Audio Soc. CD-3

Table 12 List of test sequences for the Robustness to Bit Errors test

6.4 Test Results

The following ANOVA table presents the outcomes of the experiment on robustness to degradation by the injection of bit errors. (See section 2.4, page 11, for a brief explanation of the important data analysis parameters).

Factors	df Effect	MS Effect	df Error	MS Error	F	p-level
1	1	.70438	18	1.934836	.3641	.553792
** 2	2	86.05687	36	1.049042	82.0338	.000000
** 3	6	73.78683	108	.406095	181.6983	.000000
1x2	2	.35888	36	1.049042	.3421	.712559
1x3	6	.18949	108	.406095	.4666	.831705
** 2x3	12	6.59262	216	.387274	17.0232	.000000
1x2x3	12	.40891	216	.387274	1.0559	.399164

** significant effect at p << .05

Factors: 1 = Seashore test (High, Low)
2 = Audio materials (3 items)
3 = Bit error rates (7 levels: Ref., MUSICAM No errors, MUSICAM 5 BER values)

Table 13 Robustness to Bit Errors, ANOVA Summary

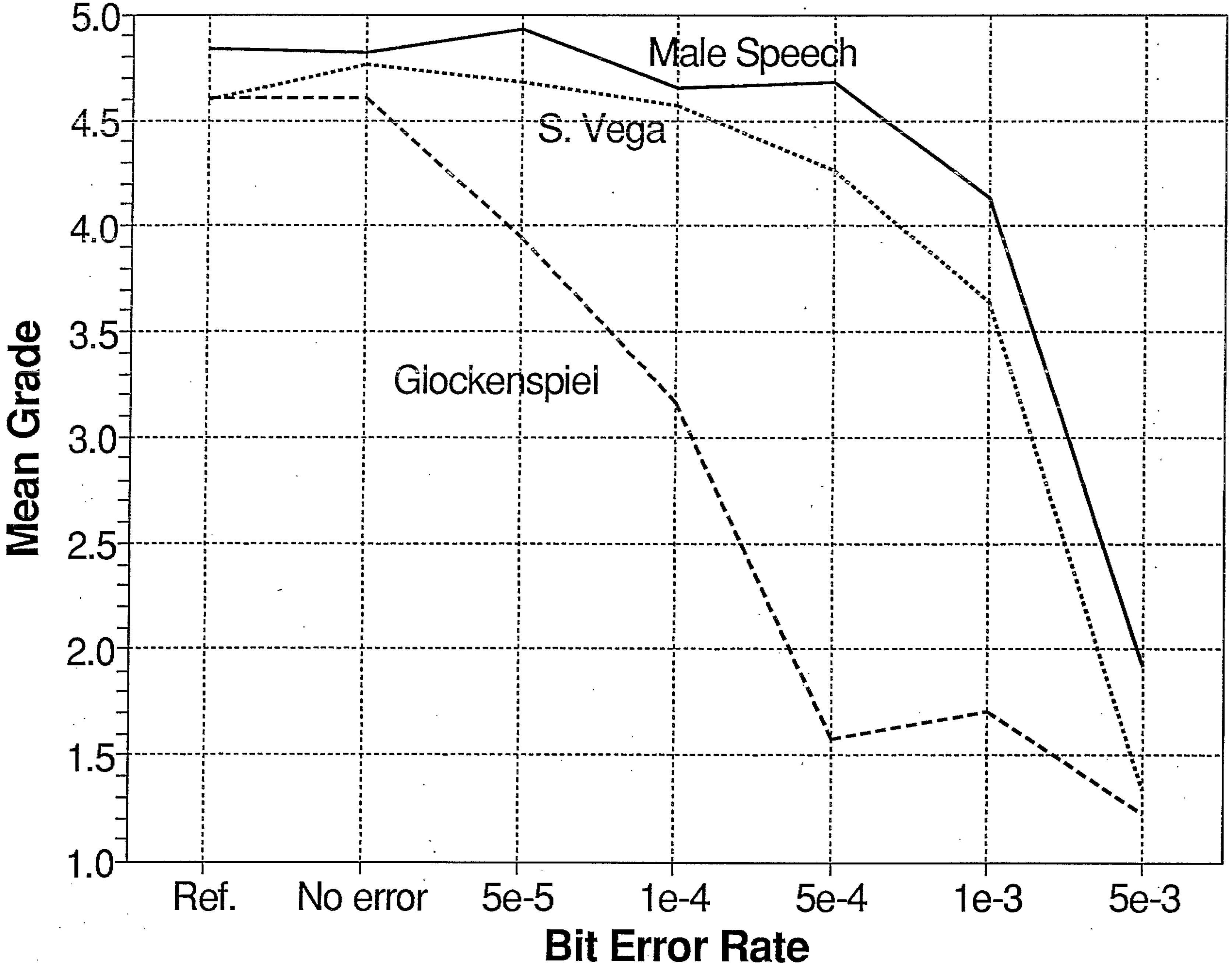
The Seashore factor played no role at all in any effects in the experiment. The other two factors, namely Audio materials and Bit errors, are involved in all three of the highly significant effects shown in the ANOVA. Under these conditions, examination of the interaction of these two factors will reveal all.the information of interest in this experiment. This interaction is presented in Figure 11.

In that figure, two of the Audio materials (Male speech and S. Vega) appear to have highly similar and roughly monotonic curvatures, and are relatively close together. The third sample (Glockenspiel) shows a different and unique pattern with a steeper overall slope and greater irregularity. It must be pointed out that, due to an erroneous manipulation in the preparation of the DAT cassette that was used in the test, the Glockenspiel signal that was recorded under the "No error" condition was a multi-pass MUSICAM signal instead of a single-pass as should have been. A mean measured rating of 3.4 was obtained for this multi-pass MUSICAM version of the Glockenspiel during the experiment. The single-pass MUSICAM version of the Glockenspiel was found to be statistically identical to the Glockenspiel reference in other experiments (Basic Audio Quality and Tandem Coding). We have good reasons to believe that a similar outcome would have been obtained here without the erroneous manipulation described above. And so, the measured rating of 3.4 has been replaced with a mean grade of 4.6 (identical to that of the Reference) for the Glockenspiel (No error) in Figure 11.

Newman-Keuls comparisons ($p < .05$) confirm that the two upper curves are indeed highly similar. In both cases, the first five points along the abscissa do not differ statistically, and the last two points are each different from each of the other six points within the curves. This tells us that for the Male speech and S. Vega materials, random errors at a rate of up to $5x10^{-4}$ did not produce any perceptible impairment. Error rates of $1x10^{-3}$ and above produced audible degradation to these two Audio samples. By contrast, the two points to the right of the "No error" in the Glockenspiel curve (i.e. $5x10^{-5}$ and $1x10^{-4}$) are each reliably different from all the other data points in that curve. The last three points are not statistically different from each other. Consequently, for the Glockenspiel stimulus, error rates as low as $5x10^{-5}$ produced audible impairment and this impairment was rated on average as "Slightly annoying". The Glockenspiel was far more sensitive to the detection of impairments due to bit errors than the other two stimuli were.

The main conclusion of this experiment is that random errors injected to a MUSICAM compressed bit stream at rates as low as $5x10^{-5}$ can produce audible degradation to some audio materials. For one test item (Glockenspiel), an error rate of $1x10^{-5}$ produced "slightly annoying" impairments whereas error rates of $1x10^{-3}$ or more were required to produce a similar kind of impairment to the other two test items used in the experiment. This conclusion is only valid however for the error protection scheme implemented in the MUSICAM system version tested in this study.

Figure 11 - Robustness to Bit Errors



7. Tandem coding capability

7.1 Purpose

The purpose of this test was to assess the subjective quality of audio material which has undergone multiple coding in cascade with the MUSICAM system. The tandem coding scenarios simulated included multiple coding with MUSICAM at 192 kbits/sec followed by multiple coding at 128 kbits/sec.

7.2 Test method

The two-stimulus (A-B) presentation along with the five-grade impairment scale described in section 3.2 was used. For each test sequence, the following combinations of A-B pairs were presented in different trials:

Combination	A	B
1	Reference	Hidden Reference
2	Reference	MUSICAM (1x128 kbits/s)
3	Reference	MUSICAM (1x192 kbits/s + 2x128 kbits/s)
4	Reference	MUSICAM (2x192 kbits/s + 2x128 kbits/s)
5	Reference	MUSICAM (4x192 kbits/s + 2x128 kbits/s)
6	Reference	MUSICAM (1x192 kbits/s + 5x128 kbits/s)
7	Reference	MUSICAM (2x192 kbits/s + 5x128 kbits/s)
8	Reference	MUSICAM (4x192 kbits/s + 5x128 kbits/s)

Table 14 A-B pairs used for Tandem Coding

The six tandem coding scenarios considered are described in column B of Table 14 above, from row 3 to row 8. The scenarios included 1, 2 or 4 coding stages at 192 kbits/s followed by 2 or 5 coding stages at 128 kbits/s. These numbers were provided by the CBC and represent typical scenarios that could occur on their network. As pointed out in section 2.1.3, <u>no conversion to analog</u> was done between coding stages.

The three test sequences described in section 7.3 were presented in the eight A-B combinations described in Table 14 to yield a total of 24 trials. The test sequences were presented in a cyclical order (seq. 1 to 3) from trial to trial but the A-B combinations varied in a manner unpredictable to the listeners. The experiment used 28 listeners and was conducted over loudspeakers.

7.3 Test material

The following subset of the ISO-IEC/MPEG test sequences were used:

Seq. #	Title	Track/Index	Time	Source
1	Glockenspiel	35/1	00:00 - 00:16	EBU SQAM 422-204-2
2	Male Speech	17/2	54:16 - 54:35	Japan Audio Soc. CD-3
3	Bass Guitar	-	20 s	RR Recording (DAT)

Table 15 List of test sequences for Tandem Coding

7.4 Test results

The following table presents the ANOVA results of the Tandem Coding experiment. (See section 2.4, page 11, for a brief explanation of the important data analysis parameters).

Factors	df Effect	MS Effect	df Error	MS Error	F	p-level
1	1	2.99756	25	3.401597	.88122	.356849
** 2	2	21.29008	50	.872519	24.40071	.000000
** 3	7	2.47767	175	.324408	7.63751	.000000
1x2	2	.11938	50	.872519	.13682	.872449
1x3	7	.33798	175	.324408	1.04184	.403695
2x3	14	.34817	350	.349724	.99555	.457185
1x2x3	14	.31534	350	.349724	.90168	.557246

** significant effect at p << .05

Factors: 1 = Seashore test (High, Low)
2 = Audio materials (3 items)
3 = Coding (8 levels: Reference, MUSICAM 1x128 kbits/s, 6 tandem scenarios)

Table 16 Tandem Coding, ANOVA Summary

Extremely reliable main effects for both Audio materials (factor 2) and Coding (factor 3) are evident in the ANOVA. The Seashore factor fell far short of significance as did all the interactions.

Both main effects are shown in Figure 12. Since all interactions were far from significance, then statistically, the three curves are parallel to each other. Looking first at the Audio materials factor, it appears that Male speech and Bass guitar curves do not differ from each other while the Glockenspiel received consistently lower ratings than both of those samples across all the levels of the Coding factor. This is fully confirmed by Newman-Keuls tests. These tests showed the difference between the Glockenspiel and each of the other two Audio materials to be highly reliable at $p < .001$, while differences between the Male speech and Bass guitar were well within chance expectation.

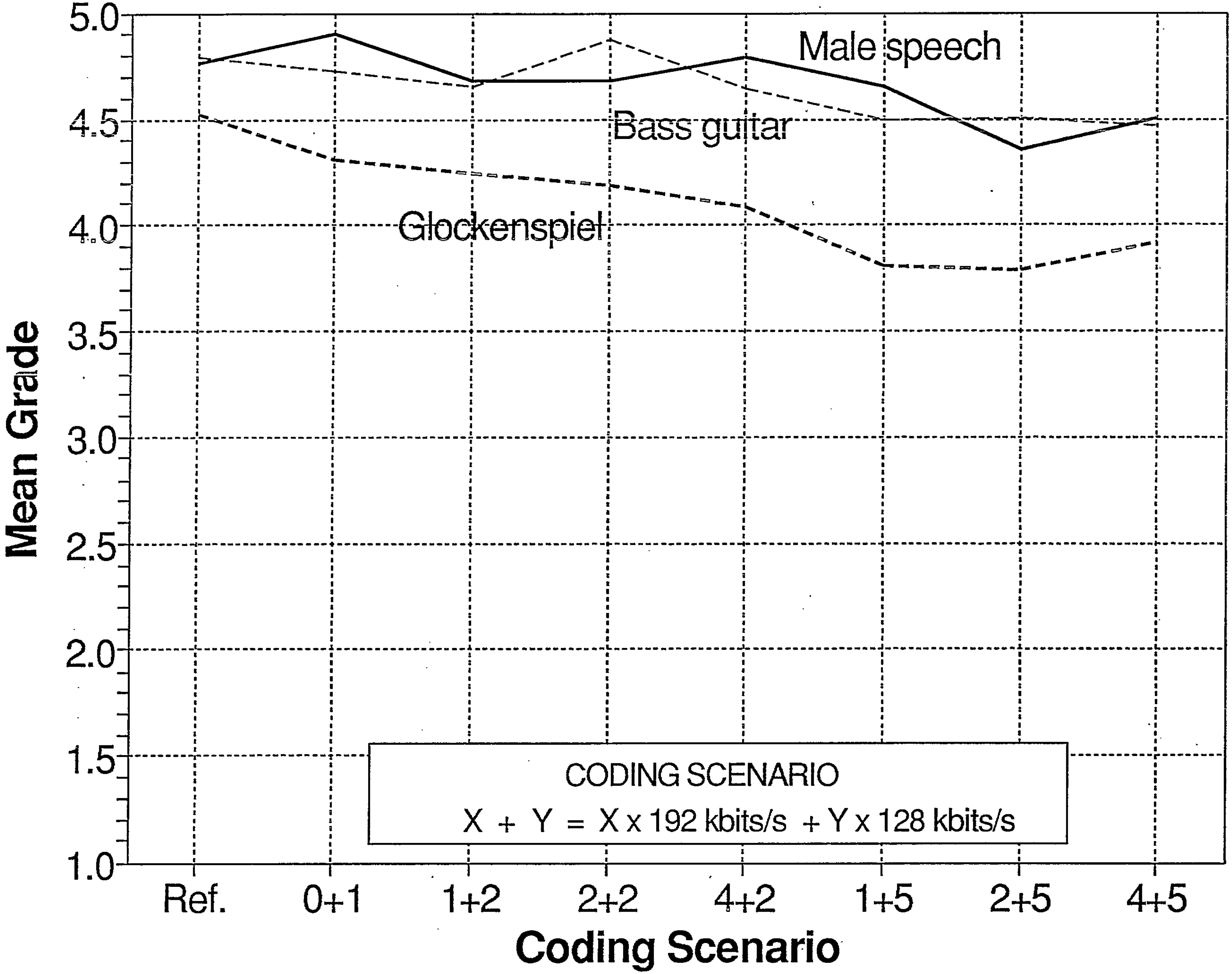
Turning now to the Coding variable, a general overall decline in ratings with increased passes is evident in Figure 12. The apparent fluctuations are statistical in nature as shown by Newman-Keuls tests. For all three stimuli, these tests place the first five levels (Reference, 0+1, 1+2, 2+2 and 4+2 into one group and the last three (1+5, 2+5 and 4+5) into a second one; the differences within the groups are attributable to chance, while the differences between the two groups are statistically reliable ($p < .05$).

These findings lead us to the following conclusions for the Tandem Coding experiment: cascading up to four coding stages at 192 kbits/s with two stages at 128 kbits/s yielded a transparent chain. From this, it can be deducted that up to four stages at 192 kbits/s alone or two stages at 128 kbits/s alone should also be transparent. However, when one to four coding stages at 192 kbits/s were combined with five stages at 128 kbits/s, the resulting chain produced a significant difference from the Reference. This difference, which is imputable to the coding stages at 128 kbits/s, was rated as "Perceptible but not annoying" for the Male speech and Bass guitar audio samples and as "Slightly annoying" for the Glockenspiel.

The "weak link" in the tandem coding scenarios investigated appears to be the number of coding stages at 128 kbits/s as one would expect. Five such stages were shown to produce "slightly annoying" impairments while two stages appeared to be transparent. The cases of three or four stages at 128 kbits/s were not investigated.

The conclusion of the Tandem Coding experiment is that up to four coding stages in tandem at 192 kbits/s or two stages at 128 kbits/s are transparent. The combination of up to four coding stages at 192 kbits/s with two stages at 128 kbits/s will also yield a transparent tandem. A cascade of five coding stages at 128 kbit/s was found to generate "slightly annoying" impairments on one audio material and a "perceptible but not annoying" degradation to the other two materials used in the experiment. The experiment did not explore the cases of 3 or 4 stages at 128 kbits/s.

Figure 12 - Tandem Coding



8. MUSICAM vs FM comparison

8.1 Purpose

The purpose of this test was to compare the basic audio quality of the MUSICAM system to that of high-quality FM. The goal was to find out if listeners had any particular preference between MUSICAM and FM. As explained in detail in section 2.1.4, high-quality FM signals were obtained by connecting an FM transmitter back-to-back with an FM receiver. No processing other than lowpass filtering to 15 kHz was performed on the analog audio signal fed to the FM encoding and transmitting equipment.

8.2 Test method

The two-stimulus (A-B) presentation along with the following seven-grade comparison scale described in CCIR Rec. 562-2 was used:

3	B much better than A
2	B better than A
1	B slightly better than A
0	B the same as A
-1	B slightly worse than A
-2	B worse than A
-3	B much worse than A

Table 17 Grading scale used for FM vs MUSICAM Comparison

A trial consisted in a presentation of two-stimuli (A-B). After each trial, the listeners was asked to grade the B sequence with reference to the A sequence using the comparison scale of Table 17. Listeners were informed that the grading scale was continuous and that they could assign score values with one decimal. Each test sequence was presented in the following 4 different combinations of A-B pairs:

Combination	A	B
1	FM	FM
2	FM	MUSICAM
3	MUSICAM	FM
4	MUSICAM	MUSICAM

Table 18 A-B combinations used in FM vs MUSICAM

The eight test sequences described in section 8.3 were presented in the four A-B combinations of Table 18 to yield a total of 32 trials. The test sequences were presented in a cyclical order (seq. 1 to 8) from trial to trial but the A-B combinations varied in a manner that was unpredictable to the listeners. The test used 25 listeners and was performed with headphones.

8.3 Test material

The following eight test sequences were used:

Seq. #	Title	Track/Index	Time	Source
1	Female speech	49/1	00:00 - 00:14	EBU SQAM 422-204-2
2	Organ solo	-	12 s	Capriccio 10040
3	Triangle	32/1	00:00 - 00:18	EBU SQAM 422-204-2
4	African song	-	14 s	Warner Broth. 925447-2
5	Piano solo	-	12 s	DG 400036-2
6	Violin solo	8/2	00:29 - 00:47	EBU SQAM 422-204-2
7	Orchestra	65/1	00:00 - 00:21	EBU SQAM 422-204-2
8	20-string koto	10/1	00:00 - 00:14	Japan Audio Soc. CD-6

Table 19 List of test sequences for FM vs MUSICAM Comparison

8.4 Test results

The following ANOVA summarizes the outcomes of this experiment. (See section 2.4, page 11, for a brief explanation of the important data analysis parameters).

Factors	df Effect	MS Effect	df Error	MS Error	F	p-level
1	1	.000343	23	2.265746	.000151	.990287
** 2	3	8.789227	69	1.295883	6.782424	.000449
3	7	.748748	161	1.081530	.692304	.678477
1x2	3	2.736927	69	1.295883	2.112017	.106605
* 1x3	7	2.931819	161	1.081530	2.710806	.011046
√ 2x3	21	1.400015	483	.898879	1.557513	.055035
1x2x3	21	1.097506	483	.898879	1.220972	.227368

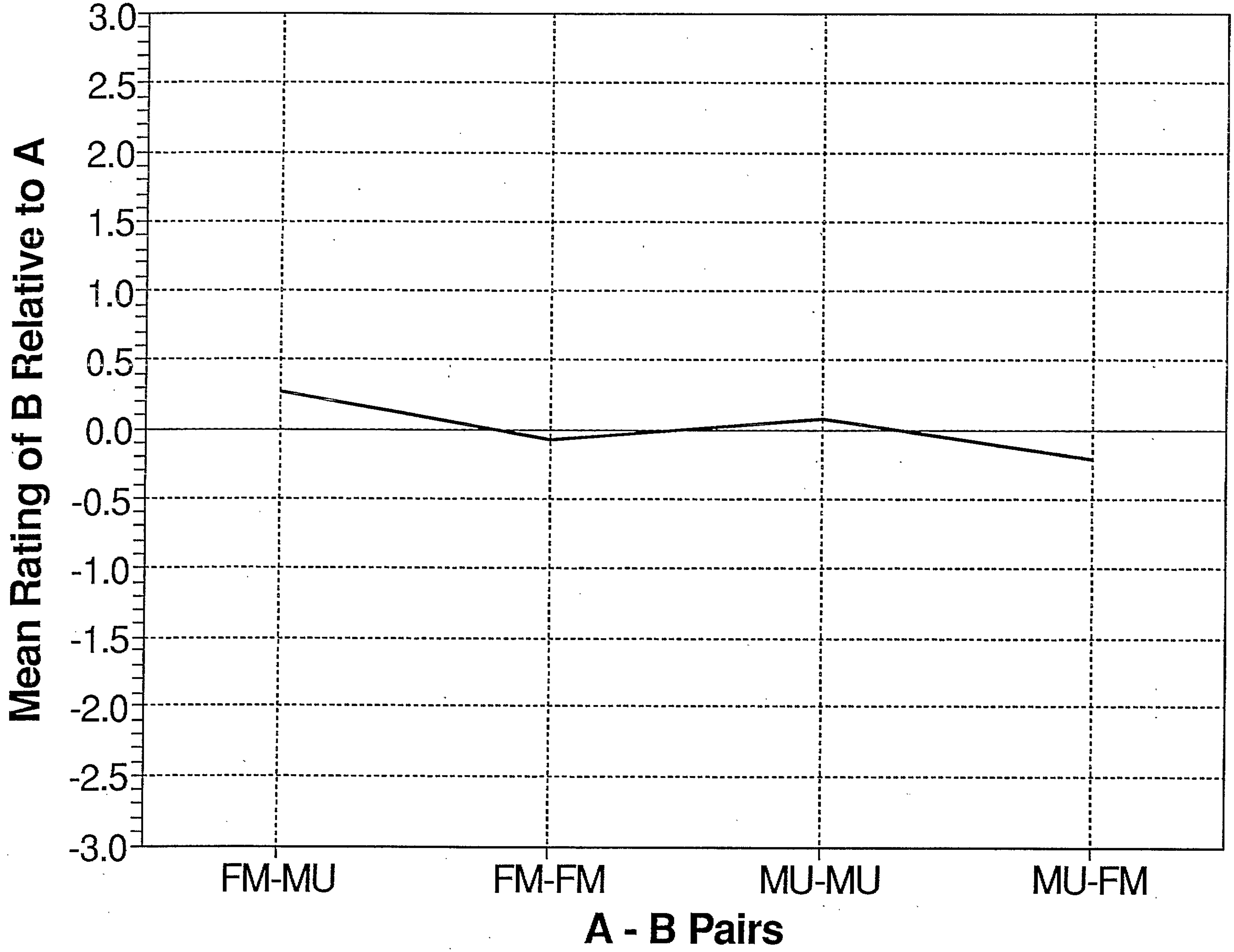
** significant effect at $p \ll .05$
* significant effect at $p < .05$
√ noteworthy effect at $p < .06$

Factors: 1 = Seashore test (High, Low)
2 = FM-MUSICAM (4 comparisons pairs)
3 = Audio materials (8 items)

Table 20 FM vs MUSICAM, ANOVA Summary

We will begin with the major finding of the experiment, the main effect of the FM-MUSICAM comparisons in which listeners rated the second member of the pair relative to the first (factor 2 in the ANOVA, Table 20). This is presented graphically in Figure 13. While the differences are very small, there is a consistent tendency for MUSICAM to be preferred in all the combinations in which a MUSICAM coded version of an Audio material is the comparison. The converse is true too, that whenever an FM version is the comparison, it is less preferred. This appears true even when FM and MUSICAM are compared to themselves. However, sub-analyses show that the FM-FM and MUSICAM-MUSICAM comparisons are not really different from each other. The FM-MUSICAM and the MUSICAM-FM ones are reliably different from each other, as well as are the FM-MUSICAM from FM-FM and the MUSICAM-FM from MUSICAM-MUSICAM differences (Newman-Keuls and Scheffe comparisons, $p < .01$ on the significant differences).

Figure 13 - FM compared to Musicam



The other statistically reliable finding concerns the interaction between the Seashore factor and the Audio materials (factor 1 by 3). This outcome, however, sheds no light at all on the major purpose of the experiment since it has nothing whatever to do with FM or with MUSICAM. It simply indicates that there were reliable differences in ratings on some of the eight Audio materials by the High Seashore versus the Low Seashore group. For interest, we report that the High Seashore group gave higher ratings to the Speech sample and to the Organ, while the Low Seashore group gave higher ratings to all the other 6 samples except for the Triangle on which there were no apparent differences as a function of Seashore level. We do not offer a sensible interpretation of the particular outcome that we got in this interaction.

Another interaction that approached but did not quite reach acceptable significance was that between Audio materials and the FM-MUSICAM comparisons (factor 2 by 3, $p < .06$). However, detailed examination of the data showed that this entire effect was due to a single material, the Organ. When the MUSICAM version of this material was compared to the FM one, a larger preference was shown for MUSICAM than was obtained with any other sample under any FM-MUSICAM condition (mean rating for that material in that condition was +1.0, $p < .01$ Newman-Keuls).

Not significant statistically but noteworthy for the trend it showed was the interaction between the Seashore factor and the FM-MUSICAM comparisons (factor 1 by 2, $p < .11$). High Seashore listeners tended to have a higher preference for MUSICAM in the FM-MUSICAM comparison and a somewhat larger negative rating for FM in the MUSICAM-FM comparisons. This again suggests that listeners who are more competent in judging musical stimuli are more discriminating in their perceptual judgments of audio stimuli. We repeat, however, that this finding is very weak and may be due to chance.

<u>MUSICAM was reliably preferred to FM although by a very small difference. The high quality FM signals used in the comparison were generated under ideal conditions which are not representative of typical FM reception by consumers.</u>

9. Conclusion

In this report, the content and the results of a series of listening tests that were carried out on the MUSICAM audio source coding system were described. The particular MUSICAM system that was tested was the version submitted in July 1990 to the ISO-IEC/MPEG committee: it performed independent coding of left and right channels of a stereo pair at a reduced bit-rate of 128 kbits/s per monophonic channel.

The following tests were performed:

1) Basic Audio Quality
2) Stereophonic Image Quality
3) Monophonic Compatibility
4) Robustness to Bit Errors
5) Tandem Coding Capability
6) FM vs MUSICAM Comparison

Two different experiments were conducted to assess the Basic Audio Quality of the MUSICAM system. The two were quite similar and differed mainly in that low anchor stimuli (i.e. deliberately impaired sequences) were used in the first experiment and not in the second one. In both experiments, listeners were unable to detect any significant differences between reference and MUSICAM encoded-decoded audio materials. Based on the listeners and experimental procedure used, the MUSICAM system tested appears to be transparent with respect to Basic Audio Quality.

In the Stereophonic Image Quality test, the binaurally recorded stimuli used produced a great deal of "mirroring" where events recorded in front of the dummy-head were often perceived at the rear by most listeners. Mirror transformations were used to reveal the systematic and symmetrical relationship between subjective and objective localizations. These transformations did not obscure the comparisons between reference materials and MUSICAM processed ones: both were perceived identically. The MUSICAM system was thus found to be transparent with respect to the Stereophonic Image.

In the Monophonic Compatibility test, stereophonic audio materials processed through the MUSICAM system were graded identically when presented in mono or in stereo. And so, the MUSICAM system appears to be compatible with monophonic reproduction. The only caution to this conclusion is that the experiment produced no differences whatever. And so, it may be argued that the failure to find any difference may have been due to insensitivity of the experiment for revealing differences. The test sequences, for instance, were perhaps not critical for this particular type of test. We temper our conclusion accordingly. Although the experiment supports monophonic compatibility for MUSICAM, additional testing is needed before a stronger conclusion can be made.

In the Robustness to Bit Errors experiment, random errors with gaussian-like distribution were injected in the MUSICAM compressed bit stream at rates ranging from $5x10^{-5}$ to $5x10^{-3}$. Error rates as low as $5x10^{-5}$ were found to produce "slightly annoying" audible degradation on one audio material (Glockenspiel). Error rates of $1x10^{-3}$ or more were necessary to produce a "slightly annoying" impairment to the other two audio materials used in the experiment. This conclusion however is only valid for the particular error protection scheme implemented in the MUSICAM system version tested.

The Tandem Coding experiment investigated the subjective quality of audio materials processed through 1 to 4 coding stages at 192 kbits/s followed by 2 or 5 coding stages at 128 kbits/s. No conversion to analog was done between coding stages. The combination of 1 to 4 stages at 192 kbits/s with 2 stages at 128 kbits/s was found to be transparent. From this it can be deducted that up to 4 stages at 192 kbits/s alone or 2 stages at 128 kbits/s alone should also be transparent. A cascade of 5 coding stages at 128 kbits/s was found to generate a "slightly annoying" impairment on one audio material (Glockenspiel) and a "perceptible but not annoying" degradation to the other two audio materials we used. The experiment did not explore cases of 3 or 4 stages at 128 kbits/s.

In the FM vs MUSICAM comparison, audio materials processed through MUSICAM (at 128 kbit/s per monophonic channel) were reliably preferred to FM although by a very small margin. The high quality FM signals used in the comparison were generated under ideal conditions which are not representative of typical FM reception by consumers.

The evidence for the usefulness of music judgement tests, such as the Seashore, in experiment of this type, is minor. Interesting but small differences between "high" and "low" scoring listeners were found only on the Basic Audio Quality experiments and in the FM vs MUSICAM comparison. "High" scorers appeared to be more critical listeners but none of the conclusions would be altered if the Seashore data were excluded as a factor in the analysis. As noted above, our listeners represented only a narrow, upper range in music judgement and so, in comparison to the general population, most were above average.

10. Acknowledgements

The authors would like to express their gratitude to the CCETT and the IRT for the loan of the MUSICAM equipment and the kind assistance of their staff: Messrs. Y.-F. Dehery (CCETT), P. Urcun (CCETT), G. Stoll (IRT), R. Sedlmeyer (IRT) and M. Link (IRT). They would also like to thank Messrs. S. Lyman, A. Duboyce and T. Caruso of the CBC for the loan of equipment, the production of FM material, their input in the elaboration of tandem coding scenarios and the recruitment of CBC staff in the listening tests.

The authors are grateful to the Swedish Broadcast Corporation for supplying a copy of the ISO-IEC/MPEG audio test sequences. They also wishes to acknowledge the assistance given by Joe Valente and Helen Vu during the preparation and the execution of the tests. Finally, a special thank you to all listeners for their invaluable participation.

References

1. G. Stoll and Y.F. Dehery, "High Quality Audio Bit-rate Reduction System Family for Different Applications", Proc. IEEE ITC '90 (1990), pp. 937-941

2. Y.F. Dehery, R. Halbert, B. Le Floch and J.B. Rault, "Digital Audio Broadcasting for Mobile Reception", Proc. of ITU-COM '89, Geneva, 1989

3. R.J. Beaton, "MUSICAM: Subjective Test Report", MPR Teltech Ltd., February 12, 1991

4. M. Vidal, "Mesure des Caractéristiques d'un Système de Transmission MF", Rapport sur Tâche no. 17-0065, Ingénierie de Radio-Canada, Décembre 1990

5. C.E. Seashore, "Psychology of Music", Dover, New York, 1967, (Republication of the book originally published by McGraw-Hill, 1938)

6. J.A. Sloboda, "The Musical Mind", Clarendon Press, Oxford, 1985

7. R. Shuter-Dyson and C. Gabriel, "The Psychology of Musical Ability", Second Edition, Methuen, London, 1981

8. MPEG Audio Test Report, Stockholm July 1990, ISO/IEC JTC1/SC2/WG11 Coding of Moving Pictures and Associated Audio

9. B.J. Winer, "Statistical Principles in Experimental Design", McGraw-Hill, New York, 1962

10. A. Gabrielsson, "Statistical Treatment of Data from Listening Tests on Sound-Reproducing Systems", Rep. TA 92, Dept. of Technical Audiology, Karolinska Inst., Sweden, 1979

11. S.E. Maxwell and H.D. Delaney, "Designing Experiments and Analysing Data", Wadsworth, Belmont, California, 1990

12. A.L. Edwards, "Experimental Design in Psychological Research", Fifth Edition, Harper & Row, New York, 1985

13. R.E. Kirk, "Experimental Design: Procedures for the Behavioral Sciences", Brooks/Cole, Belmont, California, 1968

14. W.L. Hays, "Statistics", Third Edition, Holt, Rinehart & Winston, New-York, 1981

15. R.S. Woodworth and H. Schlosberg, "Experimental Psychology", Revised Edition, Henry Holt & Co., New York, 1954

Grusec, Ted
Musicam listening
tests report.

Communications
Research Center

Shirleys Bay
3701 Carling Avenue
P.O. Box 11490
Station H
Ottawa, Ontario
K2H 8S2

Centre de recherches
sur les communications

Shirleys Bay
3701, avenue Carling
Case postale 11490
Succursale H
Ottawa (Ontario)
K2H 8S2