A SUBJECTIVE COMPARISON OF FIVE ANALOG AND DIGITAL TAPE RECORDERS

Wieslaw R. Woszczyk
McGill University
Montreal, Quebec, Canada

and

Floyd E. Toole
National Research Council
Ottawa, Ontario, Canada

Presented at the 74th Convention 1983 October 8-12 New York

AN AUDIO ENGINEERING SOCIETY PREPRINT

A SUBJECTIVE COMPARISON OF FIVE ANALOG AND DIGITAL TAPE RECORDERS

Wieslaw R. Woszczyk
McGill University, Montreal, Quebec, Canada
and
Floyd E. Toole
National Research Council, Ottawa, Ontario, Canada

ABSTRACT

Carefully controlled double-blind listening tests were performed on two popular analog studio tape recorders and three low priced digital tape recorders. Studio and concert hall performances were recorded in parallel on the machines. Twelve experienced listeners provided analytical ratings for each of five programs replayed in random sequence by the machines. Individual listeners exhibited preferences dependent on both the program and the machine. However, there was also evidence of a population preference.

0 INTRODUCTION

The introduction of digital technology in the wake of improved analog development has caused some debate between the proponents of the old and the new. Often, loosely based opinions are expressed referring to the sound quality attributes of both technologies and a truly objective point of view is rarely obtainable. Digital technology has been popularly sampled through an analog medium, the disc, and has gained an unfavorable opinion among some listeners for defects that may be not directly attributed to the digital recording process. Poor recording and editing techniques, excessive distortion allowed by inexperienced operators, faulty tape-to-disc transfer, or bad pressings could all have contributed to the criticisms of digital audio. Similarly, many professional equipment users were initially discouraged by the unreliability of digital recorders and continue to question their operational flexibility and advertised superiority in sound quality. Their opinions have been based on casual control-room comparisons which are generally unreliable but have assumed a great deal of importance.

This paper shows the results of a comparison of the subjective sound quality of several digital and analog tape recorders in fully-controlled double-blind listening tests and attempts to contribute some needed laboratory objectivity to the debate.

1.0 EXPERIMENTAL RATIONALE

It is unlikely that any single experiment will be definitive in settling an argument with as many dimensions as this one. The very fact that sound quality alone is being considered ignores many practical aspects of tape recorders such as price, size, reliability, ease of operation, editing, and so on.

If a tape recorder is regarded as simply an audio information storage device then the ultimate test of performance should be an exhaustive battery of technical tests. This, by now, has been done a number of times and the result is a display of relatively small imperfections and some questions about the long- and short-term audible consequences of the imperfections. In spite of numbers that look negligibly small, many people claim to hear important differences. In the analog-vs-digital debate strong opinions exist in both directions. From popular discussions on the subject it seems that considerations of pleasure or annoyance resulting from the reproduced sounds take precedence over the more straightforward matter of "technical" accuracy in replicating an original sound. In audiophile terms this would be interpreted as "musicality".

The experimental procedure adopted in these initial tests was selected to present the listeners with minimal restrictions in forming and expressing their opinions of sound quality. The pace of the tests was leisurely and the listening analysis emphasized neither the aesthetic nor technical aspects of the sounds. Listeners expressed their preferences in terms of the pleasurability of listening to the

reproduced sounds and in terms measured against a recollection of the original sounds. All listeners in these tests were experienced audio professionals. Eleven were professional musicians by education and recording engineers by profession. Six of the twelve listeners participated in the test-signal recording sessions the day previous to these tests, and all but one of the listeners had long-term experience of similar sessions in the same concert hall/studio complex. It is reasonable to presume, therefore, that the appraisals would be at least partially founded on a solid accumulated experience with the live "original" sounds.

The imperfections in a complex recording/reproduction chain are capable of accentuating, masking or compensating for imperfections in any link in that chain, in this case the tape recorder. This fact alone must temper the strength of any conclusions. However, the carefully controlled experimental procedure is certain to minimize the influence of perhaps one of the most powerful variables of all, the personal prejudices of the listeners. On the technical side every effort was made to employ only the highest quality recording and reproduction equipment. Three of the recordings were first-generation on each of the machines, only where necessary, in the popular music recordings, was a second generation tape used for the listening assessments.

2.0 EXPERIMENTAL METHOD

2.1 The Digital and Analog Recorders

The recorders chosen for the double-blind listening tests represent a selection of digital and analog machines popular in the audio industry. They include two analog studio machines, two mass-produced consumer-oriented digital audio recorders and one low-cost professional digital processor. The consumer digital recorders are frequently employed in professional applications and, with the exception of the bandwidth standard, they theoretically should differ from their professional counterparts only in terms of reliability and not audio performance. It was not possible to arrange the testing of limited-production professional digital recorders, except for one unit, notably a prototype.

The most obvious audible difference in performance between analog and digital recorders is that of signal-to-noise ratio and this difference must be promptly removed with the use of an appropriate noise reduction system. In preliminary tests, done without noise reduction, the listeners were easily able to identify analog recorders strictly on the basis of their tape noise. Therefore, the objective was to match the noise reduction system to the analog recorders in a way that would make all machines deliver appproximately the same subjective performance.

Figure 1 shows a summary of specifications for the five machines tested: two analog (designated as A and B), and three digital (C, D and E). Machine A was a 15ips 1/4" half-track recorder of European

manufacture, equipped with a four-band companding-type noise reduction system capable of 25 dB noise reduction, also of European manufacture. It was calibrated a for 0 VU at 250 nW/m with a 1 kHz tone and its frequency response was adjusted to be as flat as possible. Machine B was a 30 ips 1/4" half-track recorder of American manufacture combined with a four-band companding-type noise reduction system allowing up to 11 dB of noise reduction. This recorder was calibrated at +5 VU above 200 nW/m and was aligned for the flattest possible response. Machine C was a 1/2" VHS video transport integrated with a 14-bit linear PCM processor. The recording level was set with a 1kHz tone at -15 dB below 0 on the unit's meter which has a maximum reading of +5 dB. Machine D consisted of a separate 1/2" Beta video transport and a switchable 14-bit or 16-bit linear PCM processor. Both of the video transports were run in the 2-hour mode. Throughout the tests this processor worked in the 16-bit mode, chosen because this is how this unit is most frequently operated. The recording level was set with a 1 kHz tone to -15 dB on the unit's meter which has a maximum reading of 0 dB and gives an overload indication above that level. Machine E was a factory prototype of a companded-predictive-delta-modulation processor which utilised a separate 3/4" video U-matic transport. The recording level was set with a 1 kHz tone, according to factory recommendation, to 0 on the weighted volume level meter of the processor. The drop-out compensation of the video recorder was disabled for the duration of the test so it would not disturb the encoded audio signals.

2.2 Level Calibration and Noise Performance

During recording it was noticed that the 15 dB headroom at 1 kHz chosen for machines C and D was not sufficient in some circumstances and caused occasional overload. Distorted fragments of recordings were not used in the tests. High frequency pre-emphasis effectively reduced the headroom at high frequencies and it was estimated that an additional 10 dB of headroom would be necessary to accommodate all possible waveforms in all types of music. This would reduce the signal-to-noise ratio of these machines by 10 dB. Curiously, with the same overload margin for C and D, machine D indicated overload more often than C, the difference most likely being the properties of the meters. This may suggest the need for improved standardisation of level indicators.

After the level calibration, the subjective noise performance of the recorders was checked by recording with no input signal and listening to the playback of each machine. At typical monitoring levels (90 dBA) the noise performance of all machines was excellent and noise was inaudible. At maximum monitoring levels machine A sounded most quiet, followed by E, D, and C, all extremely quiet. Machine B was clearly more noisy than the others. It was felt that to make it quiet by increasing the record level above +5 VU (ref. 200 nW/m) would greatly sacrifice the distortion performance.

The subjective noise performance of the recorders was also briefly evaluated by recording tones of various frequencies. The purity of tones was most obvious on the digital machines since both analog machines showed fluctuations in level at high frequencies due to their transports. Full masking of the noise band not attenuated by the noise reduction could not be accomplished with single frequency steady-state signals. Machine E demonstrated a substantial increase in noise with tone frequencies increasing from 1 kHz to 20 kHz. Later, the spectral content of noise and the signal to noise ratio of all the machines was measured. In the preliminary tests machines C and D showed an audible increase of noise with tones around 18 kHz and 20 kHz. Very disturbing was the high number of error blips we heard on high frequency tones played back by machine D. Working in the 16-bit mode the number of blips increased when the tone frequency increased from 10 kHz to 20 kHz. Using HG (high-grade) tape did not correct the problem. However, changing the mode to 14-bit immediately removed the distortion, indicating that this machine had insufficient error correction or concealment capability in the 16-bit mode.

Throughout the three full days of listening machines C and D had four major audible drop-outs each. Machine C emitted brief "chirps" and machine D muted momentarily. During the experiments, however, listeners tended to attribute the brief mutes to careless operating by the experimenters in the control room. Machine E did not demonstrate any audible dropouts although its error correction LED would flash occasionally. Both analog machines always delivered an uninterrupted signal.

2.3 Technical Measurements

Prior to the recording, both analog machines and their respective noise reduction units were aligned for flat response and minimum distortion. The digital machines were used as delivered because the user does not have any means of adjusting the machine's response.

In addition to the subjective tests some steady-state sine-wave and impulse-FFT measurements were carried out. These measurements were done after the listening tests so that prior knowledge of machine performance could not influence opinions. The purpose of the measurements was to establish the frequency response, impulse response, signal-to-noise ratio at various frequencies, the signal spectrum (distortion) and noise spectrum, and the relative polarity change of each recorder. This objective information was gathered to aid the interpretation of the results of the subjective listening tests.

Figure 2 tabulates the response of the machines to tones of frequencies between 20 Hz and 20 kHz recorded at the calibration levels (see Figure 1). Clearly, flatness of frequency response is representative of digital recorders, although the right channel of

machine E displayed a 3 dB drop in high-frequency response.

Figure 3 presents the results of the FFT measurements of a 1 kHz tone recorded at the calibration levels on each machine. Observe the signal to noise ratio, distortion products (harmonics), and the spectral content of noise. Machine E displays an unusually poor signal-to-noise ratio (60 dB), machine D (16 bit) shows some second-harmonic content, and machine C (14 bit) has very low noise and distortion. Both analog machines A and B show very low noise and have similar distortion performance.

Figure 4 shows an FFT analysis of a 10 kHz tone for machines A, B, C, and D.

Figure 5 presents the analysis of a 20 kHz tone. Note the high-level, 24 kHz product of aliasing on machines C and D, the good noise performance of analog machines A and B, and the poor signal-to-noise ratio of machine E (55 dB) with 18 and 22 kHz sidebands.

Figure 6 illustrates the spectral content of a recording on each of the machines made with no input signal. Disregard the relative level of each noise spectrum. Note the high level of mid and low frequency noise generated by machine E.

Figure 7 shows the impulse response of each machine. Note the excellent response of both analog machines and the ringing of the digital machines C and D. The unusual time-domain response of machine E is, of course, reflected in the frequency-domain transform of this response. Machines D and E were also found to reverse the polarity of the signal, a characteristic that had been previously noted and corrected in the experimental setup.

2.4 The Sound Material, the Listeners, and the Questionnaire

The musical material was specially recorded for these listening tests since it had to fulfill many special signal requirements and had to be recorded simultaneously on all five machines. The ease of evaluating various parameters in sound recording depends largely on the characteristics of the program material. Therefore, particular care was taken to provide recordings of exceptional technical and musical quality, with a wide range of spectral and textural complexity, varying purity and coherency of spatial information, and with both steady-state and transient sound sources. The premise was that the listener should be given the opportunity to hear each of the many parameters of sound in varying contexts with different degrees of masking, in many levels of intensity and purity of presentation.

In this case five pieces of music were chosen. Two of a classical nature were recorded live in a 600 seat concert hall; music I was an oboe solo, music II a guitar duo. Music III was recorded live in the studio;

a jazz quartet with drums, bass, piano, and guitar. Music IV and V were pre-recorded "pop" pieces mixed down live from a twenty-four track recording. The original multi-track recording was done at 15ips tape speed using a 4-band companding noise reduction system of the type used with machine A. Very high quality measurement microphones were used for all of these recordings and in the recording of music I and music II no signal processing equipment was used. Care was also taken to choose good performances and good performers since there is evidence that a poor musical performance may affect the listener's ability to concentrate.

The twelve test subjects consisted of both authors, eight students of McGill's Graduate Program in Sound Recording, and two people who work in the audio field. All subjects had extensive listening experience and most had knowledge of the principles of listening analysis. All subjects passed audiometric tests confirming their normal thresholds of hearing.

The questionnaire presented in Figure 8 was used to evaluate each fragment of musical material by each listener. Throughout the three days of listening tests over 600 data sheets were collected, tabulating 5400 individual responses to each musical fragment heard in ten successive rounds. The questionnaire has been used for several years in the subjective testing of loudspeakers by the National Research Council of Canada and its usefulness has been previously established [1, 2, 3, 4].

The individual parameters (presence, brightness, fullness, softness, spaciousness, clarity, distortion/noise), in addition to any comments from the listener, provide collectively the subjective characteristics of sound quality which is then summarized by two overall ratings: pleasantness and fidelity. The pleasantness rating asks for a subjective opinion; i.e., "How much did you like this sound?" The fidelity rating requires a more objective opinion; i.e., "Do you think that this sound is true to the original"? All parameters are rated on a scale from 0 to 10.

2.5 The Organization of the Listening Tests

After the recording process was completed the outputs of all five machines were fed to line inputs of the console, each machine assigned to a separate subgroup controlling the VCA's of its individual input channels. The linearity of each console channel was previously measured and found to be orders of magnitude better than the recorders. Cables were checked for conductivity and polarity reversal. With the respective subgroup fader at maximum gain position for each machine, individual faders were adjusted on the playback of a 1 kHz tone to give exactly a 0 VU reading in the Left and Right buss of the console. A

matched pair of KEF 105 Series II loudspeakers, previously measured, was placed on individual stands about a foot off the floor. The speakers were driven by a Bryston 4B power amplifier. The LEDs indicating the reference axis were used to adjust the angle of tweeter/midrange enclosures and obtain the most extended high frequency response for all three seating positions.

Playback sound levels were adjusted to provide listeners with levels suitable for "serious" listening. At no time did the amplifier or the loudspeaker reach clipping levels (the loudspeakers had built-in protection devices).

The listening tests were double-blind by organization. The subjects were seated in the studio, visually and acoustically separated from the control room where the operator followed the switching procedure, presenting the machines in a predetermined randomized sequence. Twelve listeners were divided into four groups of three. Three seats were arranged in a line equidistant from both loudspeakers and staggered vertically just enough to enable three pairs of ears to be comfortably within the listening window of the loudspeakers. Listeners exchanged seats before hearing each of the five types of music. Each two to three minute selection of music was played in randomized sequence on each of the machines. The listeners heard each machine twice on each type of music. A preliminary round (round 0) was added to give the listeners some initial experience and thus decrease the error of the first few judgements. Switching between the machines (rounds) was not instantaneous; the sound was faded in and faded out so that the listeners would not be able to recognize the machines on the basis of their start-stop characteristics. After each fade-out a 30 second pause followed to allow the listeners to complete the questionnaire, and then the next round number was announced by the operator. This arrangement enabled the listeners to hear a large, uninterrupted fragment of music and to concentrate on analyzing the parameters of the sound and the recording. Listeners knew when the music would begin and end, therefore they could plan their listening strategy accordingly.

3.0 RESULTS

3.1 Ratings of the Individual Perceptual Dimensions

All listener ratings were transformed into numerical values and processed. Initially, the mean of the two ratings given by each listener in each perceptual dimension for each musical selection was obtained. Then the cumulative ratings of the individual perceptual dimensions obtained from 12 listeners were calculated, first separately for each musical selection, for the first-generation selections (oboe, guitar, jazz), and then for all musical selections combined. Figure 9 presents the cumulative ratings of the seven individual perceptual dimensions obtained from 12 listeners and calculated for all musical selections combined.

Careful review of the graphs suggests the following assessment of similarities and differences between the machines. In the category of clarity the listeners found few differences among machines A, B, C, and D. Machine E was perceived as slightly less clear. The quality of softness was similarly rated for machines A, B, C, and D, while machine E was perceived as somewhat more soft. In the category of fullness there was very little difference noticed among all of the machines. If anything, machine B was perceived as slightly less full than average and machine E as slightly fuller than the average. In the dimension of brightness similar ratings were given for machines A, B, C, and D, machine E was perceived as somewhat less bright than the others. Similarly for spaciousness, like ratings were given to A, B, C, and D, while machine E was perceived as being somewhat less spacious. The quality of presence was rated nearly the same on all machines. Machine B was perhaps judged to have slightly more presence, while machines D and E had slightly less presence than the others. Machines A and B were given a somewhat lower noise and distortion rating than machines C, D and E. However, it should be noted that the variance of the ratings in the category of noise and distortion is 50 to 100 percent higher than that of the other categories.

In summary, machines A, B, C, and D exhibited more similarities than differences. Machine E was perceived to have slightly less clarity, less brightnes, spaciousness, and presence than the other machines. It was judged to have more softness and fullness than the others. All of these attributes of machine E appear to be consistent with a reduced high frequency response, particularly in the more audible 10 kHz range. In addition, the variability in listener ratings was the highest for machine E in the categories of clarity, spaciousness, presence, and noise and distortion. High variability in listener ratings has previously been found in audio products which reveal significant imperfections [2].

3.2 Overall Ratings for Pleasantness and Fidelity

Again, in Figure 9, the differences between the ratings of the machines were rather small and there were large variances in the ratings. However, there persisted the same tendency for machine E to receive very slightly lower ratings than the remaining four machines and also to show the highest variance in ratings.

3.3 Discussion

Upon examining the rating results it became clear that the rating system introduced certain abberations. Individual listeners often used different starting points on the numerical scales and some used very different scale factors for indicating magnitudes of perceived differences. Even though the real differences were quite small, and

many listeners responded with appropriately small differences in scores, several listeners expanded their responses to use virtually the entire 10 point scale. Such responses tended to distort the group rating distributions, and to dominate the overall scores. Normalization of the kind used in loudspeaker evaluations [3] would be inadequate to compensate for these effects.

3.4 Rank Ordering of the Perceptual Dimensions

An expedient way of normalizing the responses, and thus decreasing distortions in the rating system, was to reduce the ratings to the essential rank ordering of preference. In this manner neither the placement of the ratings on the scales nor the scale factor used by the listener influenced the presentation of the data. Following this procedure, the subjective ratings discussed above were converted into simple rankings, one to five, based on the relative magnitudes of the mean scores produced by individual listeners. In the new assessment of data we obtained a straightforward indication of the product ranking along each perceptual dimension. Figure 10 shows the results of this data processing for each of the five musical selections and the combination of rankings for the first generation recordings (oboe, guitar and jazz) as well as for all musical selections combined. In order to assign a single numerical value to each of the cumulative ranking distributions a figure of merit was calculated as the mean of the rankings for each machine.

An examination of the ranking data provides support for the points noted in the discussion of the rating data, Section 3.1, but some of the distinctions are magnified in this presentation. It is of particular interest to note the listener responses to the first three musical selections, the first-generation recordings of unadulterated classical and jazz music, compared to the assessments using the second-generation multi-track popular music recordings.

In the clarity dimension machines A, B and C were similarly ranked, with machine D slightly lower and machine E clearly discriminated against, judging by the tendency towards lower rankings. These trends were clearly developed in the first three musical selections and further accentuated by the popular-music selections.

The ranking differences along the dimension of softness were rather small. The tendency for machine C to be ranked slightly lower in the overall softness scores could be related to the noticeably high brightness scores seen earlier. Machine E was not differentiated by listeners using the three first-generation selections but rose rapidly in softness ranking when judged on the popular music. Again this is consistent with the technical measurements and the rankings of brightness and fullness. Rankings of the remaining four machines were not sensitive to the class of recording.

Overall listener ratings of fullness indicated a fairly neutral view of machines A and B, machines C and E exhibited a tendency towards fullness and machine D a mild tendency towards thinness. With the exception of the relative fullness of machine E, there was little distinction between the other four machines on the first three musical selections. With the popular music, the distinctions became quite clear. It may be relevant that the oboe and guitar recordings had little bass frequency content, and the upright bass in the jazz recording provided sound that was stable in neither amplitude nor pitch, making sequential comparisons difficult. The regularity and spectral density of the pop music kick drum was clearly the more useful, albeit artificial, test signal for relative ratings. The fullness of machine E may well be related to the lack of brightness noted above and both could be associated with the attenuated high-frequency output seen in the measurements.

In respect of brightness there appeared to be a relatively neutral overall response to machines A and B, with machines C and D tending to be ranked rather higher. Machine E was definitely judged to be less bright on most occasions. Again this trend was clearly revealed by both classes of recordings.

Rankings of spaciousness were relatively neutral with the exception of machine E which was assessed to be less spacious in reproductions of the popular recordings. This observation is especially curious in view of the artificial nature of the spatial information in these recordings. However, the result reinforces the hypothesis that impressions of spaciousness are directly related to high-frequency response. In the case of machine E the evidence of attenuated high-frequency output and the audible consequences thereof, is very clear.

Assessments of presence using the two classes of recordings were very similar for machines A, B, C and D, and precisely opposite for machine E. Machines A, B and C were highly ranked, with machine D ranked intermediately and machine E ranked highly in renderings of the first three recordings and poorly in the popular recordings. The low ranking of machine E is perhaps connected to low rankings in such dimensions as clarity and brightness. The slight discrimination against machine D seems reasonably consistent with a similar ranking in clarity, but not with the high ranking in brightness. It would appear that steady-state frequency response is not the only factor involved in these assessments.

Rankings of noise and distortion were determined primarily by the first-generation recordings. It is appropriate that the relatively noisy and distorted popular selections should be less revealing of these differences. In general machines A and B were preferred, with machines C and D just lower in the overall ranking. Machine E again was last.

Note that for this dimension the ranking is the inverse of the rating score - lowest is best.

3.5 Overall Ranking of Pleasantness and Fidelity

It comes as no surprise that machine E received the lowest ranking scores of this group in the overall classifications assessing aesthetic and technical qualities. In general machines A, B and C received the most favourable rankings with machine D taking the intermediate position.

Combining pleasantness and fidelity scores produced the composite ranking distributions shown in the final data box of Figure 10. Examining separately the ranking resulting from first-generation classical and jazz recordings reveals the suggestion of an overall preference for the analog machines A and B, followed closely by digital machines C, D and E in that order. It is clearly a very close contest. The popular music recordings elicited some distinctive rankings, notably an unambiguous preference for digital machine C and a moderately strong critical comment on digital machines D and E.

Pooling the data from all recordings reveals very much the same picture that we have seen developing throughout these presentations of data. Machines A, B and C are apparently very competitive, and all highly regarded, with machine C having a slight advantage due to its superior ratings on the popular-music selections. Machines D and E are not all that far behind but, due to an accumulation of small criticisms they have fallen subtly lower in overall performance. Machine D, in particular, was downrated mainly because of its lower ratings on popular music.

4.0 DISCUSSION

When all the results are reviewed, there are clearly no dramatic differences of opinion from this evaluation of digital and analog recorders. The minor distinctions noted by the listeners find some confirmation in the technical measurements. Machine E, which obtained the lowest total scores, displayed a deviant impulse response in the technical measurements. Its frequency response at 10 kHz showed a 1 and 3 dB roll-off, while the other recorders had flatter responses. However, a 4.0 to 4.5 dB roll-off in the response of machine A at 20 kHz was not identified by the listeners, probably because there is very little audio energy in that frequency range. Machine E also demonstrated more noise than the other machines and the distribution of its rankings in the noise and distortion category indicated that some listeners responded accordingly. It should be pointed out that the listeners were not able to consistently identify the extended low-frequency response of the digital machines. For example, the fullness rankings of machine D are not better than those for machine B which had 18 dB of roll-off at 20 Hz. This may be explained because extreme low frequencies are not present in all of the musical selections and in any

event are imperfectly reproduced by even a good loudspeaker in the listening room.

Another factor to consider is that the domains of presence, brightness, fullness, clarity, spaciousness, and softness are not independent of each other. They seem to be related to a common root, the frequency response, and the ratings in these categories can partially be explained by the frequency responses demonstrated by the machines in the technical measurements. For example, machine E, which obtained a relatively low score in the brightness category, was also judged to be slightly less present and less clear, while its fullness and softness were rated high. However, machine D, which received a slightly higher brightness rating, was also given slightly lower presence and clarity ratings. In this case, the frequency response alone does not provide a complete explanation, and other factors must be considered.

It is further evident that the different program selections produce variations in ratings depending, presumably, on the extent to which different signals reveal the virtues and limitations of the machines. This is to be expected since, in music signals, specific problems may be masked by other portions of the signal. Ideally, test signals could presumably be contrived that would concentrate the full signal energy or waveform parameter on specific technical imperfections characteristic of tape recorders.

Perhaps moving in the direction of such contrived signals are the popular music selections used here. These performances have little basis in reality since the generous use of equalizers, time delays and other signal processing has destroyed any such direct connection. Nevertheless, it is clear from the results presented here that this music was indeed useful and, for the most part, served to reinforce or accentuate the results from the classical and jazz selections. A notable exception to this was in the category of noise and distortion where it was of little use because of its inherent contamination with these same factors. On the other hand, the high dynamic intensity, repetitive nature, and widespread spectral information appeared to provide a number of important listening reference points.

This observation raises an important question about the fundamental basis of listening assessments. If such artificial signals are useful in revealing the presence, if not the precise magnitude, of imperfections it means that much of the rating is being done on the basis of which sound is least objectionable rather than which sound is more true to some original. This observation has been noted in earlier loudspeaker assessments [2] and it would appear to be true here as well.

In several respects these listening tests are tests of the test itself. The basic procedures used here have been applied extensively in the field of loudspeaker evaluation, where they were adequate to reliably discriminate differences between even very good loudspeakers. However, the differences here are measurably and audibly much smaller and it is not surprising that the data do not demonstrate convincing statistical

significance. There are, however, clear suggestions of trends in ratings and rankings that are logically related to each other and to certain clear technical differences. It is reasonable to presume, therefore, that improvements of the experimental method would provide more definite indications of the relative merits of these machines and the reasons for the discriminations. Even so, the fact would remain that the differences existing here are extremely small by comparison to diffferences between, for example, loudspeakers.

5.0 CONCLUSIONS

Blind listening tests are well known for producing controversial results. The results of these tests follow the familiar pattern. There were, from various points of view, unexpected similarities and unexpected differences. It is probably safe to say that, taking a general overview, the machines performed in a manner that inspired rather than destroyed confidence in the technologies represented; there were large similarities and small differences.

Listeners individually had their favored machines, in fact every machine had some proponents, but they were not outstandingly reliable in their identification of the favorite, or any other machine. In fact, the clearest result shows the identification of a prototype machine that had easily-measurable technical faults. Even that result is more of a suggestion of a trend than a clear-cut rejection.

Analog and digital machines were not readily distinguished. Some of our very experienced listeners tried, from time to time, to guess which machine they were listening to. Most failed. Some of the trends and preferences indicated in the results appear to be explicable in terms of straightforward measurable defects, yet there are others that thus far defy explanation. None seem to be convincingly correlated with the analog or digital nature of the machine.

The results presented in this paper apply to the consequences of a single, carefully executed, recording/reproduction process. It is suspected that the evaluation of succeeding generations of recordings might present us with different sets of data.

6.0 REFERENCES

1. A. Gabrielsson, "Dimension Analysis of Perceived Sound Quality of Sound Reproducing Systems", J. Acoust. Soc. Am., Vol. 65, pp. 1019-1033 (1979 April)

2. F.E. Toole, "Listening Tests - Turning Opinion Into Fact", J. Audio Eng. Soc., Vol. 30, pp. 431-445 (1982 June)

3. F.E. Toole, "Subjective Measurements of Loudspeaker Sound Quality", presented at the 72nd Convention of the Audio Engineering Society,

Anaheim, Oct. 1982, Preprint No. 1900.

4. "Listening Tests on Loudspeakers", International Electrotechnical Commission, Publ. 268-13: Sound System Equipment, pt. 13; in press.

7.0 ACKNOWLEGEMENTS

The authors wish to acknowledge the assistance of David Kelln, John Hill and David Findlay in the preparation and running of the experiments. We wish also to thank the representatives and the manufacturers who were most cooperative in loaning two of the machines tested here.

	ANALOG RECORDERS		DIGITAL RECORDERS			
	A	B	C	D	E	
TAPE FORMAT	1/4 INCH 1/2 TRACK 15 IPS(38CM/S)	1/4 INCH 1/2 TRACK 30 IPS(76CM/S)	INTEGRATED 1/2 INCH VHS VIDEO CASSETTE	SEPARATE 1/2 INCH BETA VIDEO CASSETTE	SEPARATE 3/4 INCH U-MATIC VIDEO CASSETTE	TAPE FORMAT
CALIBRATION	0 VU REF 250 NW/M	-5 VU REF 200 NW/M	-15 DB REFERRED TO METER ZERO	-15 DB REFERRED TO METER ZERO	0 WEIGHTED VOLUME LEVEL	CALIBRATION
NOISE REDUCTION	4-BAND COMPANDER 25 DB NR	4-BAND COMPANDER 11 DB NR	14-BIT LINEAR PCM	14 OR 16-BIT LINEAR PCM	COMPANDED PREDICTIVE DELTA MODULATION (CPDM)	DIGITAL FORMAT

FIGURE 1 THE SUMMARY OF SPECIFICATIONS FOR THE FIVE MACHINES TESTED.

MACHINE	A		B		C		D		E	
FREQUENCY	L	R	L	R	L	R	L	R	L	R
20 Hz	-4.0	-4.5	-18.0	-18.0	0	0	0	0	-0.75	-0.5
30	+1.5	+1.0	-7.0	-7.0	0	0	0	0	-0.5	-0.5
60	+1.0	+1.0	+0.5	+1.0	0	0	-0.15	-0.15	-0.5	-0.25
100	0	0	0	+0.5	0	0	0	0	0	0
1.0K	0	0	0	0	0	0	0	0	0	0
10.0K	0	0	0	0	+0.5	+0.5	+0.5	+0.5	-1.0	-3.0
15.0K	-0.75	-0.75	0	0	0	+0.25	+0.25	+0.25	-0.5	-3.0
18.0K	-2.0	-2.0	-0.5	-0.5	0	0	0	0	-0.25	-3.0
20.0K	-4.5	-4.0	-1.0	-1.0	-1.0	-0.75	0	-0.25	-0.5	-3.0

FIGURE 2. THE RESPONSE OF THE FIVE TESTED MACHINES TO TONES OF FREQUENCIES BETWEEN 20 Hz AND 20 KHz.

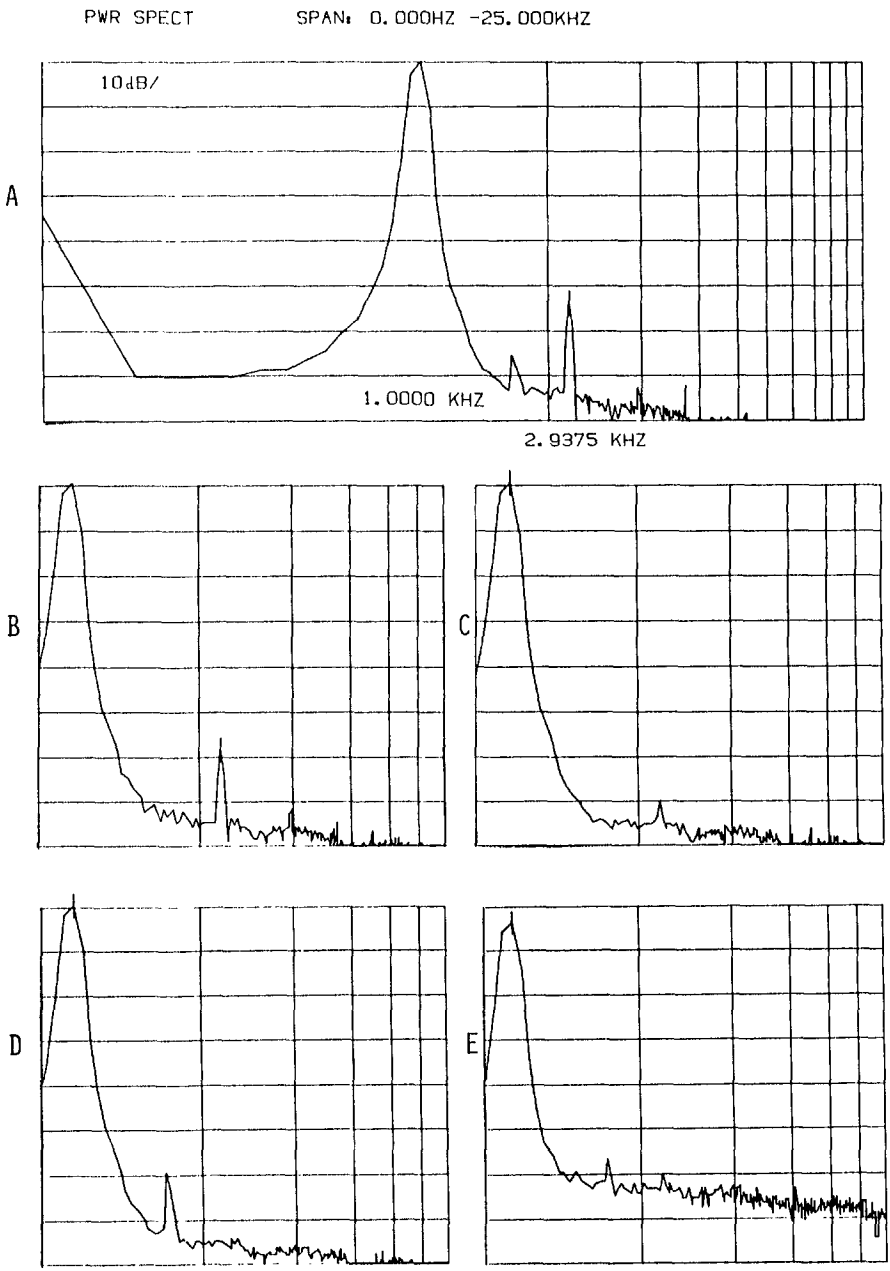

FIGURE 3. THE FFT MEASUREMENTS OF A 1 KHz TONE RECORDED AND REPRODUCED ON EACH OF THE MACHINES. VERTICAL SCALE 10 dB/DIV. HORIZONTAL SCALE 2.5KHz/DIV.

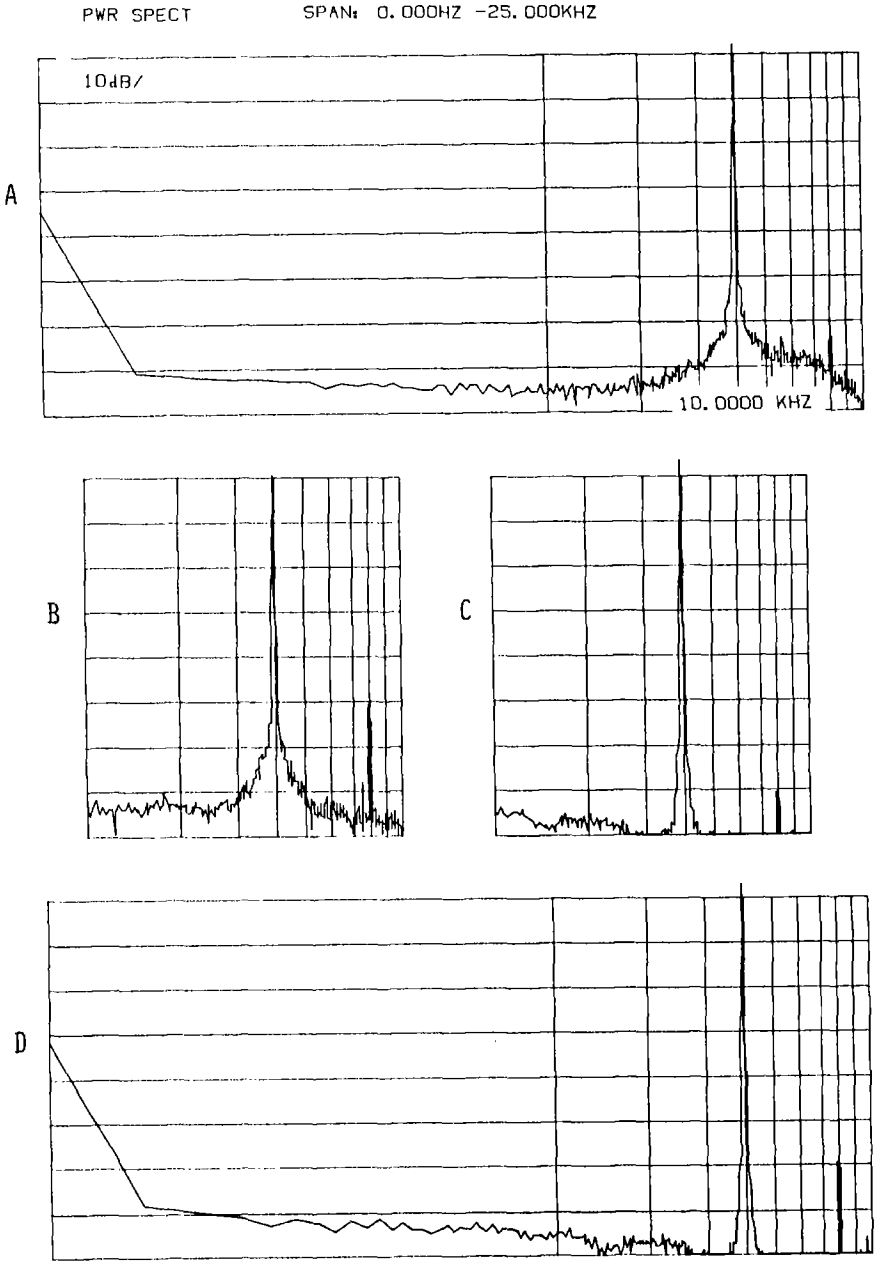

FIGURE 4. THE FFT MEASUREMENTS OF A 10 KHz TONE RECORDED AND REPRODUCED ON MACHINES A, B, C, AND D. VERTICAL SCALE 10 dB/DIV. HORIZONTAL SCALE 2.5 KHz/DIV.

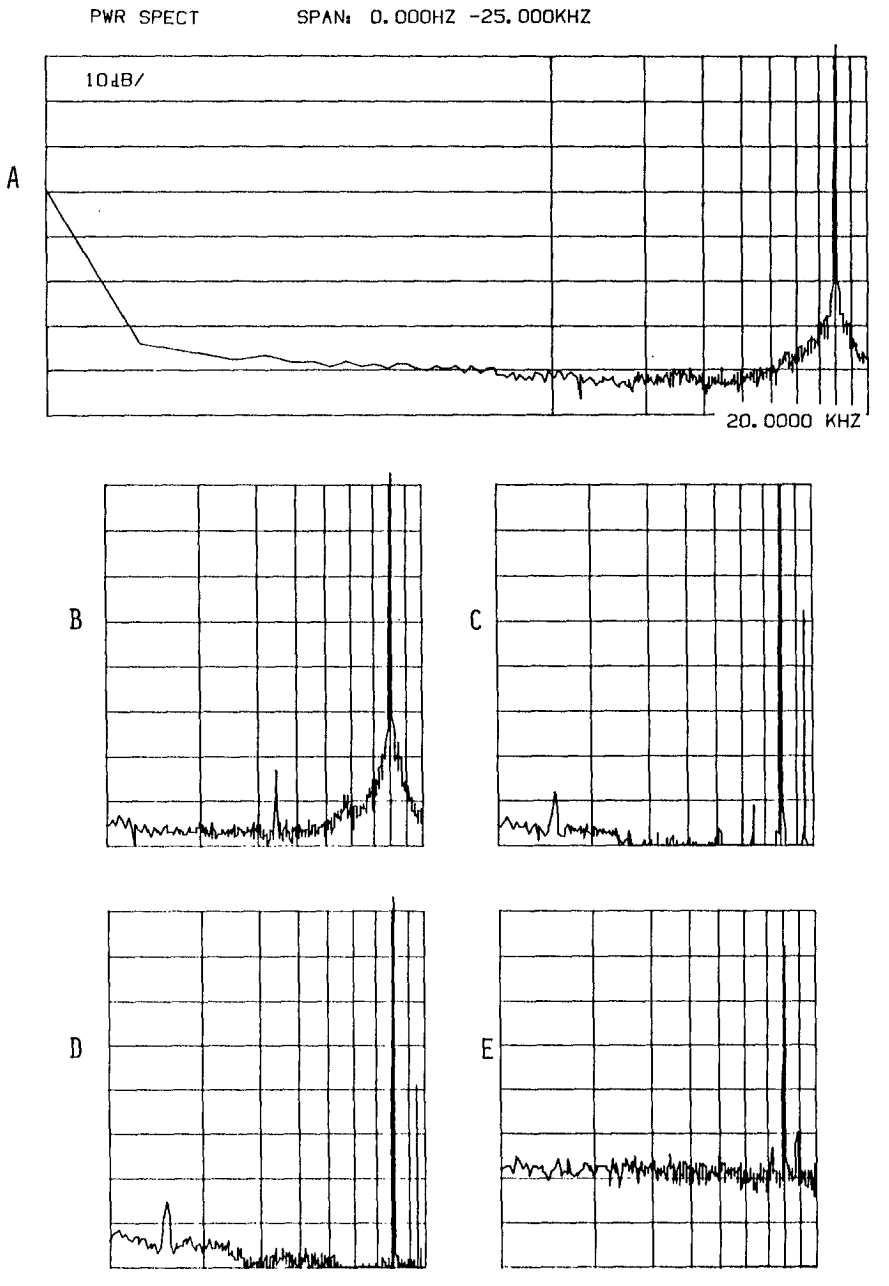

FIGURE 5. THE FFT MEASUREMENTS OF A 20 KHz TONE RECORDED AND REPRODUCED ON EACH OF THE MACHINES. VERTICAL SCALE 10 dB/DIV. HORIZONTAL SCALE 2.5 KHz/DIV.

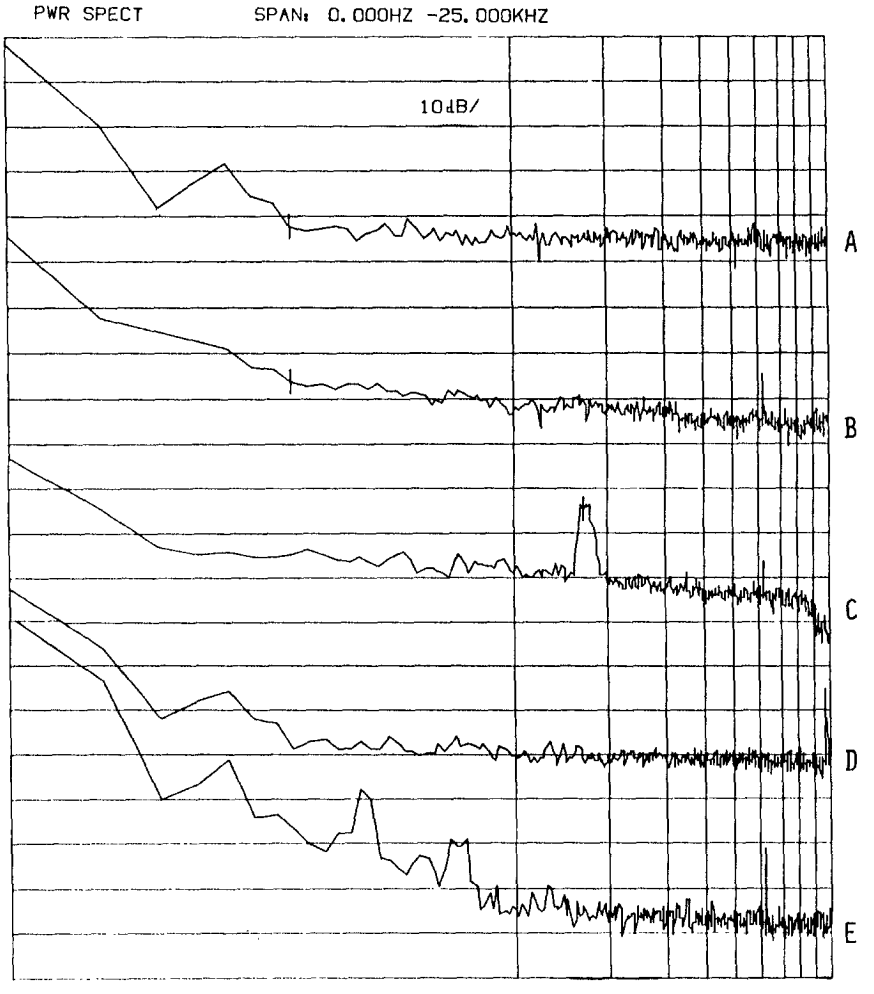

FIGURE 6. THE SPECTRAL CONTENT OF NOISE REPRODUCED BY EACH OF THE MACHINES FROM RECORDINGS MADE WITH NO INPUT SIGNAL. THE VERTICAL SCALE SHOULD BE USED TO INTERPRET THE SPECTRUM SHAPE NOT THE COMPARATIVE LEVELS BETWEEN MACHINES. VERTICAL SCALE 10 dB/DIV. HORIZONTAL SCALE 2.5 KHz/DIV.

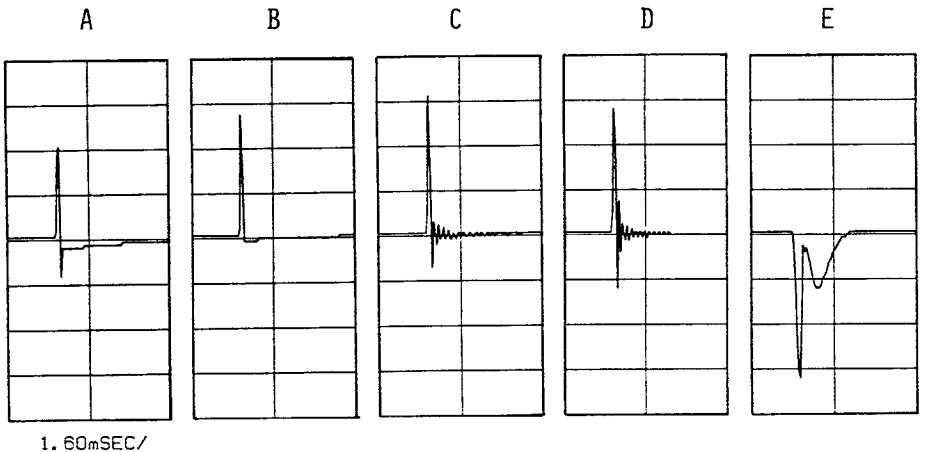

FIGURE 7 THE IMPULSE RESPONSE OF THE FIVE MACHINES. THE IMPULSE RESPONSE OF MACHINE D IS SHOWN INVERTED.

NAME		SPEAKER NO.	
DATE	ROUND NO.		
SEAT NO.			

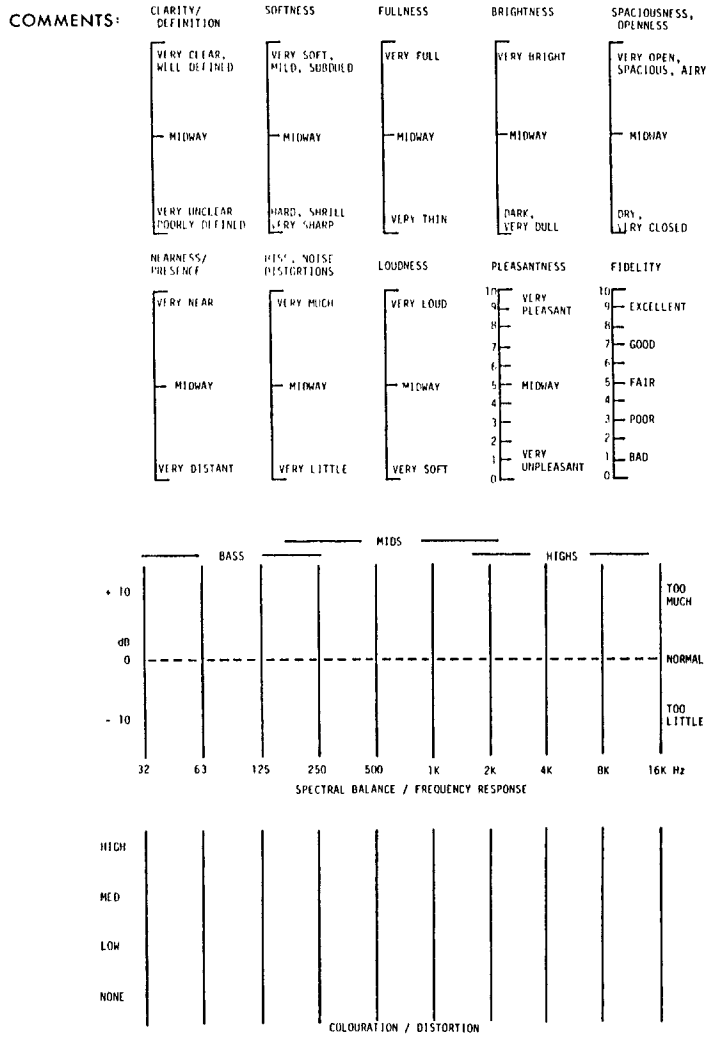

FIGURE 8. THE QUESTIONNAIRE USED TO EVALUATE EACH MUSICAL FRAGMENT BY EACH LISTENER. IN TOTAL, 5400 INDIVIDUAL RESPONSES WERE COLLECTED. THE LOUDNESS CATEGORY WAS NOT USED.

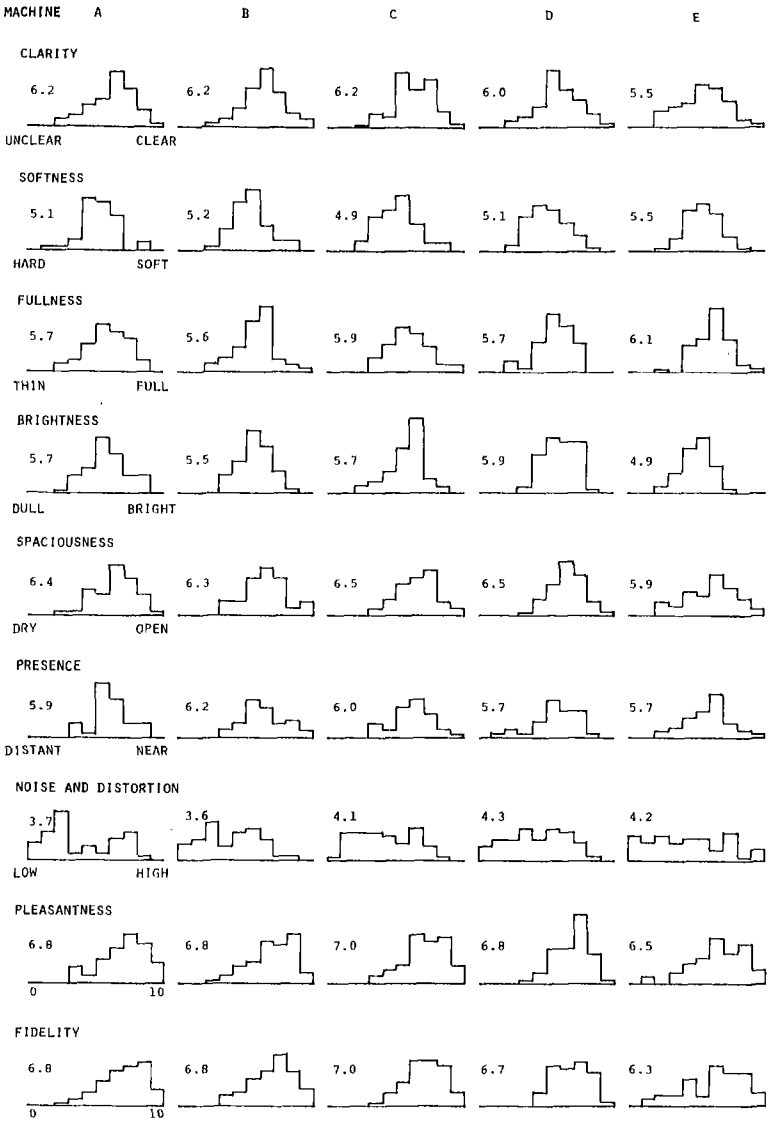

FIGURE 9. CUMULATIVE RATINGS IN THE SEVEN PERCEPTUAL DIMENSIONS AND THE TWO OVERALL CATEGORIES, OBTAINED FROM ALL TWELVE LISTENERS AND INCLUDING RESPONSES FOR ALL FIVE MUSICAL SELECTIONS. THE BASE OF EACH HISTOGRAM , FROM LEFT TO RIGHT, IS THE SAME SCALE THAT, FROM BOTTOM TO TOP, EXISTED ON THE LISTENER QUESTIONNAIRES. THE NUMBER ADJACENT TO EACH HISTOGRAM IS THE MEAN OF THE DISTRIBUTION, CALCULATED ON A SCALE OF TEN, AS IN THE PLEASANTNESS AND FIDELITY CLASSIFICATIONS.

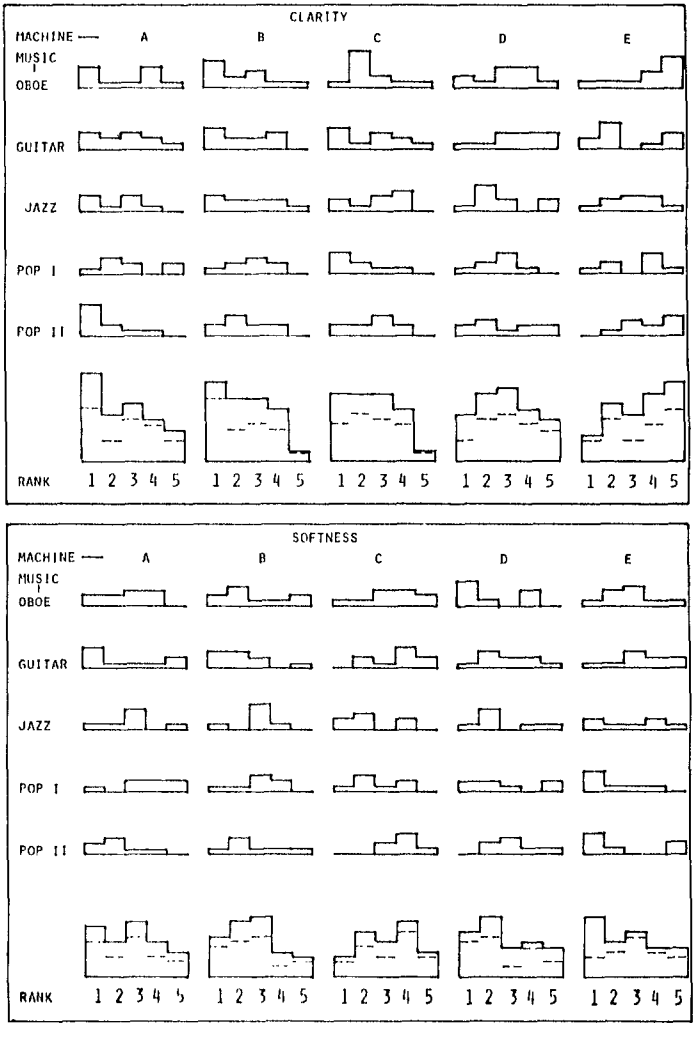

FIGURE 10. CUMULATIVE RANKING DISTRIBUTIONS DERIVED FROM THE ORIGINAL RATINGS. DISTRIBUTIONS ARE SHOWN FOR EACH MUSICAL SELECTION, FOR THE COMBINATION OF ALL SELECTIONS (SOLID-LINE HISTOGRAM), FOR THE COMBINATION OF THE FIRST THREE SELECTIONS (BROKEN-LINE HISTOGRAM) AND FOR THE LAST TWO SELECTIONS (THE DIFFERENCE BETWEEN THE SOLID-AND BROKEN-LINE HISTOGRAMS).

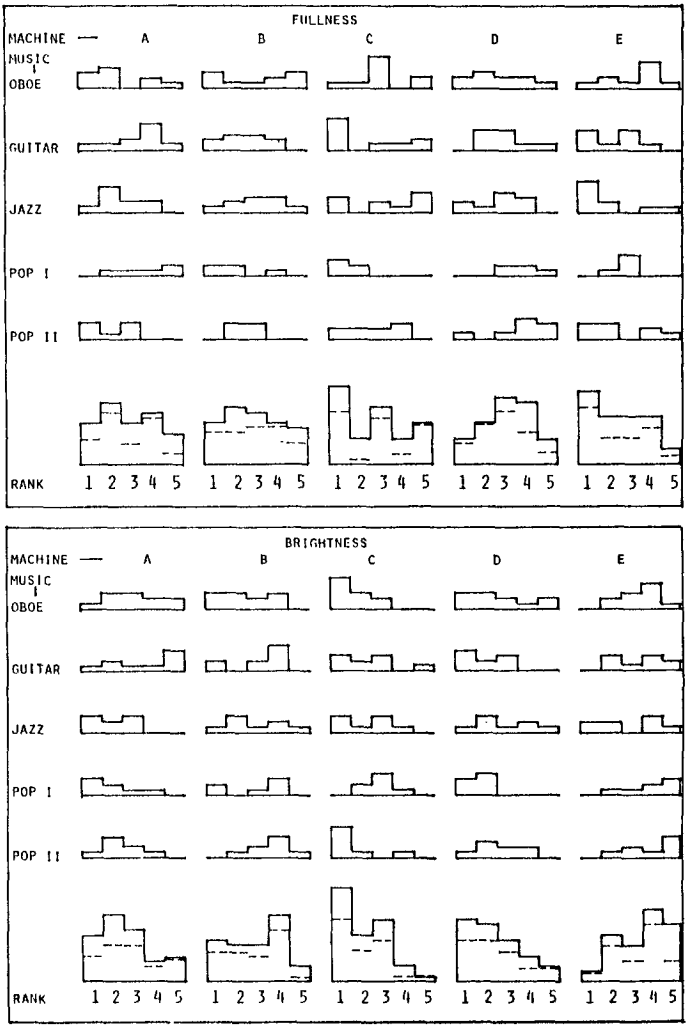

FIGURE 10. CONTINUED

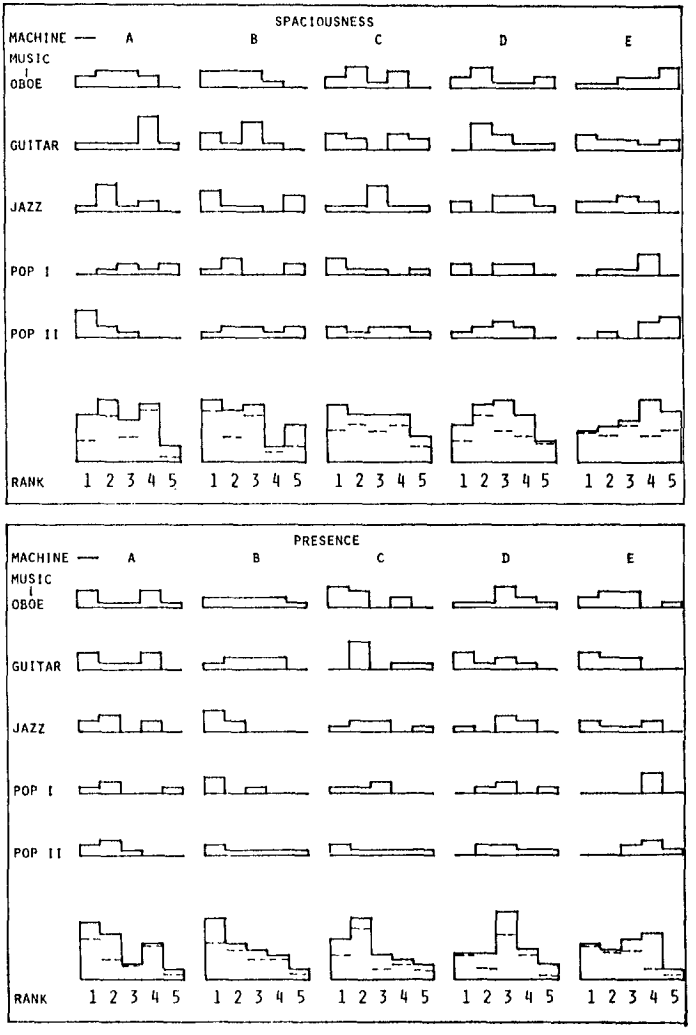

FIGURE 10. CONTINUED

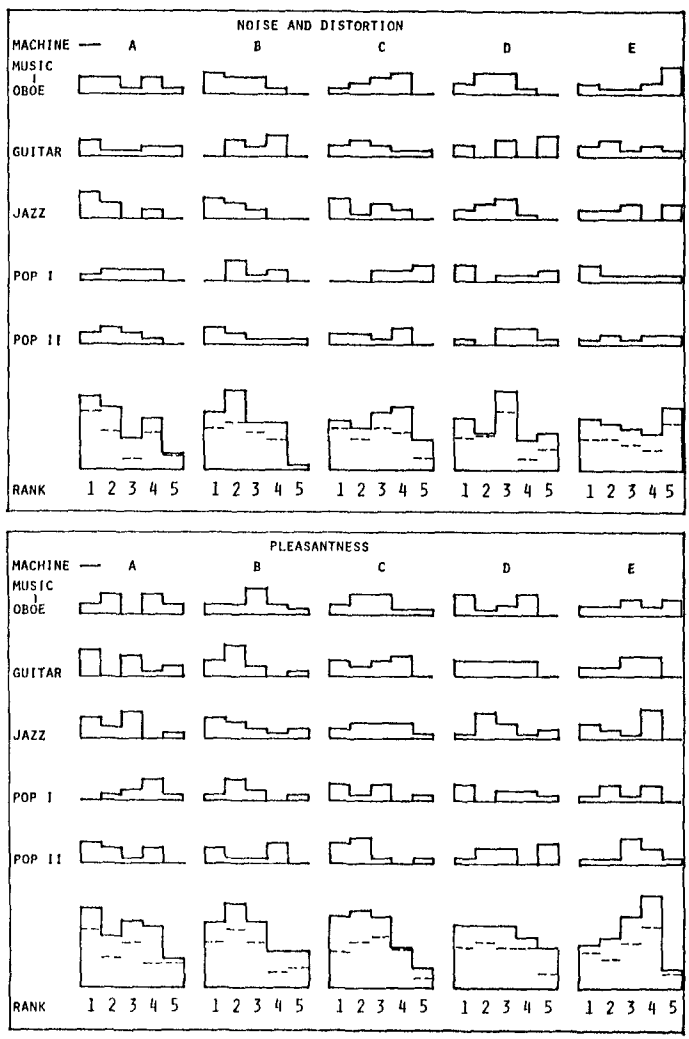

FIGURE 10. CONTINUED

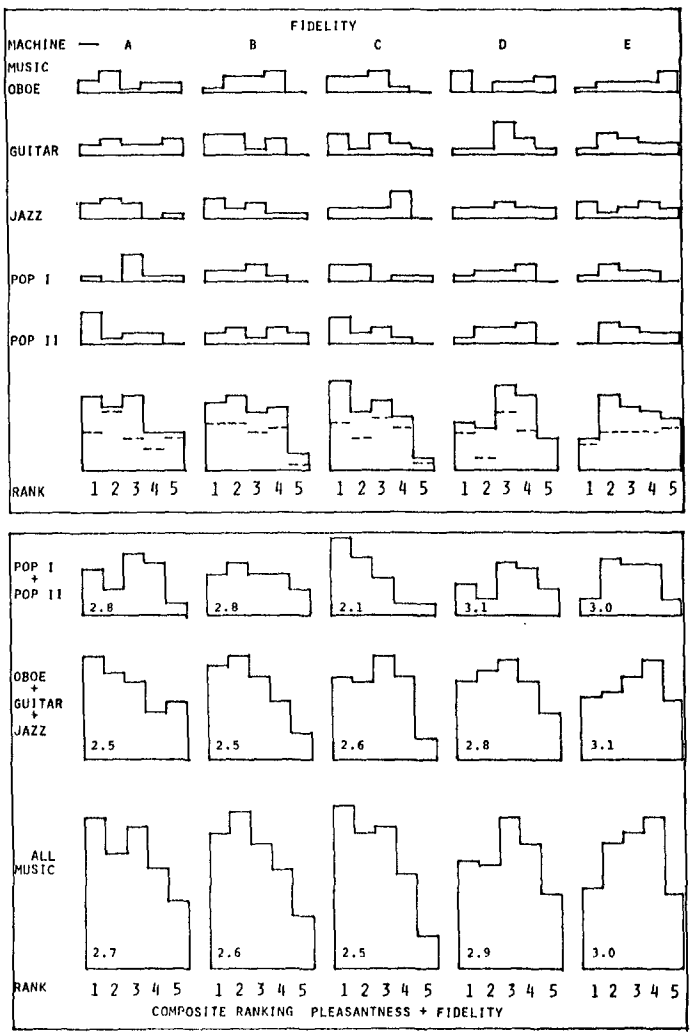

FIGURE 10. CONTINUED