THE INSTITUTION OF ELECTRONIC AND RADIO ENGINEERS

Third International Conference on

TELEVISION MEASUREMENTS

(BROADCASTING AND DISTRIBUTION)

Thursday 18 to Saturday 20 June 1987

The Casino, Montreux, Switzerland

Publication No. 74

TELEVISION MEASUREMENTS

(BROADCASTING AND DISTRIBUTION)

The Casino, Montreux, Switzerland

18th — 20th June, 1987

PublicationNo. 74

INSTITUTION OF ELECTRONIC AND RADIO ENGINEERS

ORGANISING COMMITTEE

Chairman: G. A. McKenzie, BSc, CEng, FIERE
(Consultant)

Members: D. J. Bradshaw, BSc(Eng), AMIEE
(Designs & Equipment Dept, BBC)

R. R. Ferbrache, BSc, CEng, MIERE
(Tektronix Ltd)

I. F. Macdiarmid, CEng, MIEE
(Consultant)

B. W. Osborne, MSc, FInstP, CPhys, CEng, FIEE, FIERE
(British Cable Services)

J. D. Tucker, CEng, FIERE, MIEE
(John Drew Tucker Associates Ltd)

D. E. O. N. Waddington, CEng, FIERE
(Consultant)

L. E. Weaver, BSc, CEng, MIEE
(Consultant)

D. Willats, BSc, CEng, FIERE
(Marconi Instruments Ltd)

Secretary: R. Larry, CEng, MIEE, FIERE

ERRATUM

Incorrect page sequencing
Page 84 to follow pages 85 and 86.

Published by
The Institution of Electronic and Radio Engineers
99 Gower Street, London WC1E 6AZ

ISBN 0 903748 72 X

Printed in Great Britain by Alderman Printing & Bookbinding Co Ltd
Russell Road, Ipswich.

Third International Conference on

TELEVISION MEASUREMENTS
(BROADCASTING AND DISTRIBUTION)

CONTENTS

Papers are listed as closely as possible in their order of presentation at the conference.

Session 1: Subjective Measurement

Session 2: Philosophy of Measurement

Session 3: Operational (1)

Session 4: Operational (2)

Session 5: Audio

Session 6: RF Measurements

Session 7: MAC Transmission Measurements

SUBJECTIVE ASSESSMENT OF TELEVISION: A PERSONAL VIEW

I. F. Macdiarmid*

Introduction

Subjective assessment of picture quality (or sound quality) is the basic method of judging the performance of any television (or sound) system. Objective measurements are of limited use unless they can be related to subjecative quality. Although the details of subjective test methods used in different laboratories may differ, and some of the differences have been debated for may years, it is believed that many of the differences are more aparent than real and like is not being compared with like. An attempt is made in this paper to highlight some features of subjective tests which, if adopted generally, could lead to more uniform results from different laboratories.

It should not be overlooked that subjective tests may have widely different objectives and consequently they may require different test methods. This paper is directed primarily towards laboratory tests made to assess the performance of a system with varying degrees of impairment introduced by signal processing and/or transmission.

Observers

Because of the different perceptions of picture quality by different observers and the variability of assessment by each observer, a reasonably large number of observers is required to establish with some degree of confidence a mean value for a test condition. Type of observer, ie "expert" or "non-expert", is often a matter of debate. Which should be used should depend only on the objectives of the test but, in practice, it tends to depend on the type and size of the establishment conducting the tests. The expert television specialist forms a very tiny proportion of the population at large and so, in a subjective test aimed at determining a standard for the population at large, his opinions should play a negligible part.

---

* Consultant

Grading scales

Grading scales are used in most subjective assessments and there has been much debate in quality the same between "fair" and "good" as it is between "poor" and "fair"? Examination of the results of tests made with several different scales suggests that observers tend to divide the scale up uniformly paying little attention to the exact words associated with each grade.

Adaption

The adaption of the observers to the range of picture qualitites to be assessed is of viatal importance to the standardation of results. This process is also known as anchoring. The simplest method is to ensure that, in an experiment, all grades of the scale are used and the grand mean score is close (within about $\pm 4\%$) to the mid-point of the scale. The display of identified unimpaired pictures has been proposed as a means of improving anchoring.

Opinion distributions

For the presentation of results, mean scores and standard deviations are most frequently used. The standard deviation is most useful when applied to distributions approximating to the Gaussian shape but in subjective tests using 5-step grading scales, only the opinion distribution at the centre of the scale is symmetrical and there the standard deviation is a maximum. As the extremes of the scale are approached the distribution becomes increasingly skewed and the standard deviation decreases, becoming zero at grades 1 and 5. A stepped distribution based on the binomial expansion provides a more useful model of the distributions obtained with quality or impairment grading experiments.

Presentation of results

Results presented in the form of a plot of mean scores against an objective measure of an impairment provide a number of points which may be connecated by straight lines. The use of a suitable mathematical model for the relationship will permit the disjointed graph to be replaced by a smooth curve. Not only is this useful in permitting interpolation between the data points but also it smooths the effects of the

statistical sampling. The logistic function is a suitable model which is easy to use. Its use is simplified by the use of a unit called the "imp" awhich has been regarded with suspicion by some people. The "imp" (short for impairment unit) is merely an algebraic transform of the mean score, ie a chage of variable. In fitting the logistic function to the data points (transformed into imps) three parameters must be adjusted. These are the mid-opinion value of the impairment, a slope parameter and a parameter which corresponds to the asymptotic value of the mean score as the added impairment is reduced to zero (this parameter is sometimes described as the residual impairment). Some residual impairment exists in nearly all test results and experimenters have used different methods of correcting for it - three methods are outlined in CCIR Rep 405. When the impairment is small there is probably little difference between the results from each method but the method which uses it as a parameter in fitting the logistic function to the data seems to be the most reliable.

In many applications the use of curve-fitting programs on a computer is unnecessary. The experimental data can be converted from mean scores into imps, the residual impairment determined from the score for unimpaired pictures and subtracted from the impaired data points. The corrected data points are then plotted as log(imps) against log(impairment). These points should lie on a straight line which, as a first approximation, can be drawn by eye with a ruler. A better approsimation is easily and quickly available using a linear regression program on a calculator. The straight line so obtained is called an "impairment characterisitc" whose parameters "mid-opinion value" and "slope" together with the value of the residual impairment define the results of the subjective experiment. The familiar S-shaped curve of mean score versus impairment (in log units) can be easily obtained from the impairment characteristic when required.

Conclusion

It is believed that the greater part of the differences between results from different laboratories could be avoided if sufficient attention is given to the choice of observers, the grand mean score of the experiment, the correction for resigual impairment and the use of the ogistic function model to express the results.

SUBJECTIVE ASSESSMENTS — GOOD LUCK OR GOOD PLANNING?

D. Wood* and B. L. Jones**

SUMMARY

Subjective assessments are an essential, and unavoidable, part of the development of new systems. They are also the basis by which objective parameters are used for operational performance monitoring. The paper reviews some of the difficulties associated with them and suggests what seem to be the best recently developed methods intended for the assessments of small impairments. The two most promising methods are outlined, and the paper looks forward to studies of the possible use of an alternative, a ratio-scaling method, to circumvent certain shortcomings. Finally, possible approaches for assessments of HDTV are briefly considered.

1 Introduction

The science of subjective assessments has many facets. Subjective picture grading is a regular operational practice for the broadcasting community. This includes the monitoring of the performance of the various links which make up the broadcasting chain, and the assessment of the suitability of source material for transmission. In addition, subjective assessments are an indispensable tool in the research and development of new systems and services.

The objective parameters used to assess the performance of systems, such as K-ratings, only have any value if and when there is a predictable relationship between these commodities and picture quality. At some point, it would have been necessary to assess and agree the relationship between picture quality and these objective parameters, by subjective assessment methods. In the end, what matters about image systems is subjective picture quality. We need good, reliable, subjective assessment methods, because everything else is built on these foundations.

It would be difficult to say that the science of subjective assessments is currently complete. Over the last few years, a period of convergence of ideas on scales has turned into a period of divergence of ideas on methodology. The more knowledge is advanced, the more alternatives seem to open up.

In general however, subjective assessments are a subject that many broadcast engineers prefer to shy away from. Recently, a very well respected UK researcher said that he had a filing cabinet full of names and tendencies in his office, and that by judicious selection of assessors, he could achieve any mean result we wished in picture assessments. Unfortunately, the nettle of this world of uncertainty and indeterminacy has to be grasped, because subjective assessments cannot be avoided. The following are recent examples of this.

In 1980 and 1981, a large number of laboratories in many parts of the world took part in the study of a digital conventional television standard. At the end of the day, the subjective assessments results formed an integral part of the 4:2:2-standard decision-making process. Today, we are about to embark on the standardization of transmission systems derived from the 4:2:2 production standard, and so there is a clear and urgent need for a common methodology which can be used to evaluate systems fairly, on a common basis in different parts of the world.

The study of high-definition television is the major preoccupation of many organizations in broadcasting today. The work, which hopefully will lead to common world-wide HDTV standards, has already required many subjective assessments. One of the key determinants of the optimum HDTV field rate was thought to be the quality of conversion from HDTV to conventional television, and a method was needed which would allow repeatable assessments to be made in about ten different laboratories throughout the world. The results established a common ranking order, but showed a lack of absolute consistency which is still being debated.

In the last CCIR Study Period, an Interim Working Party set up by the CCIR Study Group 11, IWP 11/4, assembled a set of reference curves (ref.1) of impairment magnitudes versus objective parameters (Rec. [AR/11]). The final curves are as complete as the IWP were able to make them, and should be used a guide for operational practice in future.

* European Broadcasting Union, Chairman of CCIR IWP 11/4.
** Vice-Chairman of CCIR IWP 11/4.

However, such curves are only as good as the input data, and when data obtained by different experiments is not the same, the only recourse is to average it. There was certainly a degree of dispersion in the source data used in these curves. This comment does not detract from the great value of the Recommendation; it simply highlights the need for a concentrated effort to unify methodology.

CCIR Interim Working Party 11/4 is now charged with trying to rationalize the existing subjective assessment texts, and with examining methods which may be relevant for tomorrow's world of high-definition and enhanced definition television. This means taking a hard look at the subject, and trying to agree methods which are relevant, accurate and reproducible. Whether this is possible remains to be established, but we have to start somewhere. The purpose of this paper is to explain the starting points as they are perceived by the Chairman and Vice-Chairman of 11/4. Time will tell if they receive support of the IWP and can be substantiated.

2 Fundamental issues in methodology

Why is it apparently so difficult to obtain high inter-laboratory correlation of results?

The first facets to examine are certain fundamental problems, which may mean that whatever precise methodology we use, results will not be consistent in different parts of the world.

In the past, a large number of grading scales were developed for subjective assessments, but at the current time (ref.2), only a comparison scale and two direct evaluation scales are recommended by the CCIR. The two evaluation scales are the five-grade quality scale (excellent, good, fair, poor, bad) and the five-grade impairment scale (imperceptible, perceptible but not annoying, slightly annoying, annoying, very annoying). These are widely used for operational monitoring and system evaluation.

2.1 The conception of the meanings of the grading terms, and the equality of the intervals between grades

In making subjective assessments, we trust that there is a popular conception of the meaning of 'good', 'excellent', etc., in the public at large, which is similar. This is not unreasonable, but how realistic is it to say that the conception of the meanings of the five CCIR quality grade terms are the same in English and, for example, Japanese; and indeed, across the range of the world's languages? If they are not, how can we fairly make and compare subjective assessment results in different parts of the world?

A first step in the analysis of this question is a report (ref.3) of an examination in the English and Italian languages of potential and used quality-scale descriptors. The results suggest that there are serious differences in conception in the English and Italian equivalents. There are differences amounting to effectively 1/2 grade or more in the conceptions of the equivalent terms, as they are currently used. This suggests that there is a real and urgent need to investigate the situation world-wide, and to do something about it.

Another facet revealed by these results was that the conception of the intervals between the currently-used quality scale descriptors, either in English or Italian is not uniform. The interval in English between excellent and good, and between poor and bad, is much smaller than between good and fair, and fair and poor. In spite of this, much of the processing done on subjective assessment results is done on the assumption that the intervals betweeen grades are equal. When we see a report that "the loss in quality was a half grade", what value is this unless we know precisely where the loss occurred in the non-uniform grading scale continuum?

Another study referred to in ref.3 of the intervals associated with the CCIR Impairment Scale concluded that, for this scale, the intervals are essentially equal between the descriptors. This may come a a relief to the supporters of the impairment scale. But there are also those who argue that the impairment scale is inherently defective, because there is conceptually no gap at all between grades 4 and 5 (either you can see an impairment or you cannot), and that the gaps between the other grades are finite because we are now considering degrees of annoyance, which is a continuum.

However this may be, there is a clear need to clarify the conceptions of the meanings of the descriptors used, for both the quality and impairment scales, and the intervals between them, in those different parts of the world wherever subjective assessment results are likely to arise.

2.2 The choice of material for the assessments

If an operational engineer is grading a transmission, he has no problem at all with the choice of material. The problem comes when we have to choose which test material to use to evaluate a system's performance. The guidelines offered by CCIR Recommendation 500-2 are that the material should be 'critical but not unduly so' for the impairment in question. As those who have made assessments will know, interpreting this in practice is not at all simple.

We might imagine that if we could hypothetically pass an infinite number of programme hours through the system to be evaluated, and evaluate somehow the quality

at each of an infinite series of small time units, we could arrive at a probability distribution with respect to the degree of criticalness of programme material. This might be, for example, a normal distribution. Where on this distribution would 'critical but not unduly so' fall? On the falling and more critical part, certainly, but precisely where is not clear. This kind of notion also illustrates the need for care when averaging the results from a number of test pictures, unless the results are either all the same or are, by chance or design, evenly spread about the hypothetical criticalness distribution.

To illustrate that the greatest care is needed in selecting test material, it can be noted that in tests of bandwidth versus quality, the same quality grade is possible with a 20% difference in bandwidth, by judicious choice of test picture (ref.4).

2.3 The choice of assessors

In laboratory tests, the question always arises as to the type of assessors which should be used. Should they be experts or non-experts; and where exactly does a non-expert stop and an expert begin?

It seems to the authors that the objective of assessments of quality is to obtain an approximation, which is as near as possible to the mean grade which would be given by an infinitely large number of assessors; that is, to the general public. Intuitively, it seems likely that non-expert observers will be closer to the population at large than experts. Of course, this is not the same as choosing people who do not really understand what they are supposed to do in the tests, who are easily distracted, unreliable, and who not have good visual acuity and normal colour vision. All these things are important if a small number of assessors is to be used as an approximation to the public.

One usually particularly useful group (we find) is young people (e.g. students) between about fifteen and twenty-five years of age. It would be wrong to say this is a recommendation, but it is certainly worth considering.

Having considered some general grey areas for assessment methodology, we still have to do the best we can in the circumstances. The authors believe that this currently amounts to using the methodologies described in the following sections.

3 Methods currently available for the assessment of small impairments

One of the most pressing needs today is for a completely specified methodology which will allow the assessment of small impairments; that is to say systems which have inherently very good quality and where there is need to judge fine differences and fine distinctions. This represents the requirement for the evaluations of bit-rate reduction systems currently being made by the CCIR.

Such high accuracy methodology is likely to be elaborate, but if it also works over the whole quality range, it could form the basis of a single test method (or a small number of test methods), which would embrace all requirements.

Several years ago, the EBU introduced into the CCIR a proposal for a high-stability method (ref.5). The **EBU method** uses the CCIR impairment scale and an implicit reference. The basic unit of the test involves the presentation to a group of five assessors of, first, a reference unimpaired picture (the assessors are told that this is so) for 10 seconds, followed by a period of grey, then 10 seconds of the test picture. After this, assessors are asked to vote on the second picture. The method includes recommendations that unimpaired pictures should also be included in the test pictures, that the grade means score should be arranged to be close to mid-opinion (grade 3), that the overall duration of the test should be limited, and other features.

It has been argued that impairment grades 4 and 5 are particularly easy for assessors to interpret. Does the picture look identical to the reference? If yes, the answer is grade 5. If something can be seen, but only just, the answer is grade 4. Consequently, high stability can be expected in this area. It has also been suggested that a mean grade of 4.5 is a convenient (although arguably arbitrary) assessment of a 'just perceptible' point, where half the people will see something and half will not.

The method is widely used, but it certainly has some drawbacks. They are, firstly, that the method does not take account of any system effects that may actually enhance the appearance of the picture, such as sharpening effects, and secondly, that the achievement of an overall mean grade of 3 often requires the artifical insertion of test material (with similar types of impairment), and this is not always easy to arrange.

Subsequent to the EBU method being proposed, an alternative high-stability method, the **double-stimulus continuous quality scale** (ref.6) was put forward to the CCIR. This method also presents assessors with pictures in pairs. One is the unimpaired reference and the other is the test picture, but the assessors are not told which is which. They are simply asked to grade each one. Assessors use two continuous lines, rather like thermometers, which are each subdivided into five sub-sections corresponding to one of the CCIR quality scales. The use of a continuous scale improves the precision with

which assessors can express their opinion, and the inclusion of an unimpaired reference, although not explicit, also helps precision. In the first variant of this system put foward, an assessor worked individually, and had personal control of the switching of the pair of pictures, such that he could switch backward and forward, until he was satisfied that he could grade them appropriately. With the assessment procedure used in this way, it also seemed that assessors were less sensitive to context; that is to say, it was not necessary to have the full range of impairments included in the assessments. This was a very promising feature, since it is sometimes inconvenient to arrange for the full range of (similar) impairments to be available.

In a second variant of the method, groups of observers were used and they are presented several times with each of the pair of pictures by outside control. This approach is certainly more efficient, but there is, arguably, insufficient evidence to be sure it is more immune to context effects than other methods. It was employed for a major series of digital codec assessments (ref.7), and the results were widely used. Something however which seemed surprising to the authors in these tests was that non-experts apparently gave more critical assessments than experts. This may be because experts gave lower grades to the references, and thus the differences were smaller, but this is not certain.

In general, the authors believe that the two best methods currently available for subjective assessments of small impairments are the above described EBU method and the double-stimulus continuous quality scale method. Both of these methods are thought to have drawbacks. The problem is that the alternatives seem to be even less satisfactory. These two methods can also be used for assessments where the major interest is lower quality, below mid-opinion, but there the precision available is unlikely to be substantially better than with single stimulus methods.

It is possible to address the question of which of the two methods is more useful by asking if there are relatively well definable occasions when the quality scale is the more relevant, and other occasions when the impairment scale is the more relevant, and we certainly hope to do this in CCIR IWP 11/4. However, we are also coming to the following view. Equipment development is extremely expensive. There can be no argument that subjective assessments often represent the crowning part of the study, so it is not unreasonable to suggest that assessments should be performed with both of the above methods where possible. This would provide a useful back-up, extra information, and what is more, will help to further an understanding of the science of subjective assessments.

4 A possible future reappraisal of assessment methods

One approach being studied in the CCIR IWP 11/4, which may circumvent many of the drawbacks of the conventional approaches, is the use of ratio-scaling. In a sense, it is a return to fundamentals.

In some recent studies in the subject, a method (ref.8) was adopted whereby assessors were asked to assign any number they felt comfortable with to the quality associated with a given test picture. They were then asked to do the same for each of the sequence of test pictures that followed. In each case, the number they gave should be proportional to the quality as perceived. That is, for example, double the size, for a quality which is considered twice as good. In this way, the assessments proceeded. At the end of the session, the assessors were asked to record which number they would assign to an 'ideal' picture, if they were to be presented with one. Following this, in the processing of the results, the ratings given to the 'ideal' pictures were brought together, and the ratings for all the test pictures were then scaled together to be comparable.

First reactions might be that a method such as this would be difficult for Mr. and Mrs. Public to follow, as assessors. However, students certainly mastered the method, and produced results which were less context sensitive than results achieved by the conventional quality scale. It remains for this method to be evaluated in other laboratories, but it certainly is worth doing. If it proved stable and generally usable, it could conceivably become the single method for the evaluations of new systems in world-wide studies.
In the area of new systems, a key subject of interest is systems which have a quality higher than conventional quality, such as enhanced television and high-definition television.

5 Assessment of high-definition television and enhanced television

People in the broadcasting community are familiar with operational monitoring practice, and many have a well trained eye for picture quality. As a consequence of this, there is a popular conception of what, for example, grade 5 means. It could be for example an RGB picture from a conventional camera with a well lit studio set. How, therefore, do you convey a measure, to this community, of the improvement associated with a high-definition picture, which might have double the vertical and horizontal resolution?

One approach could be to 'extend' the grade range, perhaps beyond five to grades six or more. In Japan, many tests were done (ref.9)

with a grading scale that included a grade beyond excellent, "ideal", and they have certainly proved that it therefore can be done. But the results published by the Japanese suggest that the area beyond excellent was in practice relatively little used.

How then can we equate HDTV quality with our existing results about quality? The authors believe that this may well be an unrealistic expectation. Conventional subjective tests are performed at a given viewing distance, for example, 6H. At this distance, the potential resolution of a 625/50 full bandwidth picture is close to the eye's limiting resolution.

So at 6H, the extra definition provided by HDTV is, in a sense, almost surplus to requirements, and it would be unreasonable to expect a dramatic increase in an assessor's grading when HDTV pictures are seen. HDTV is essentially designed to provide the same limiting resolution at a much closer viewing distance, i.e. 2.5H - 3H. At this distance, we can expect a conventional picture to be marked down considerably when assessed at the same time as HDTV, and HDTV to receive, potentially, grade 5.

In general then, the main difference between HDTV and conventional television assessments can be solely viewing distance, and there seems no reason at the moment to make any major changes to methodology apart from this. It will, arguably, never be possible to simply translate all our existing results into the new context of HDTV. But then, at least the CCIR will be kept busy for many years to come.

6 Conclusions

Subjective assessment methodology has many different facets, and it has not been possible to consider them all in this paper. The processing of results is a major area of importance, and includes the possibilities for impairment magnitude addition and the assessment of data reliability. However, before the processing can be done, we have to be on firm ground with the raw data. This paper has examined some of the uncertainties of obtaining raw data for subjective measurements. These include the choice of material, assessors and scale-usage. Although not free from drawbacks, the paper suggests what currently seem to be the two most stable assessment methods: the double stimulus continuous quality scale method and the EBU method. There is good reason to study, as a future option, a return to the ratio scaling techniques. Finally, the paper suggests that in the study of systems offering higher resolution than conventional systems, it may simply be appropriate to adapt the existing methodologies to a shorter viewing distance.

7 Acknowledgements

The opinions expressed in this paper are those of the authors, but the paper reports studies and discussions by members of CCIR IWP 11/4 who were nominated by many Administrations, and whose work is gratefully acknowledged.

8 References

1 CCIR Recommendation [AR/11], CCIR XVIth Plenary Assembly, Dubrovnik, 1986.
2 CCIR Recommendation 500-2 [MOD F], CCIR XVIth Plenary Assembly, Dubrovnik, 1986.
3 Jones, B.L. and McManus, P.R. 'Graphic Scaling of Qualitative Terms', SMPTE Journal, November 1986.
4 CCIR Report 959 [MOD F], CCIR XVIth Plenary Assembly, Dubrovnik, 1986.
5 Bernath, K., Kretz, F. and Wood, D., 'The EBU method for organizing subjective tests of television picture quality', EBU Review - Technical, no. 192, 1981.
6 Allnatt, J.W. and White, T.A., 'Double-stimulus quality rating method for television digital codecs', Electron. Lett., 16, pp. 714-715, 1980.
7 Macdiarmid, I.F. and Darby P.J., 'Double-stimulus assessment of television picture quality, EBU Review - Technical, no. 192, 1982.
8 Jones, B.L. and Marks, L.E., 'Picture Quality Assessment: A comparison of Ratio and Ordinal Scales', SMPTE Journal, December 1985.
9 'High Definition Television', NHK Technical Monograph No. 32, June 1982.

PHYSIOLOGICAL MEASUREMENTS AND DEVICES IN DIGITAL SOUND

H. N. Teodorescu*, S. Lozneanu†, E. Sofron§, L. Bucholtzer§ and A. Stoica§

SUMMARY

Digital sound allows higher quality easier measurements of the complex sound parameters describing the physiological perception. The paper is based on this statement.

The background of the problem is exposed, analysing such physiological nonlinear acoustics phenomens as complex masking, time frequency dependent time and frequency resolution, phase discrimination, and so on. These parameters of the hearing system are translated into measurable sound parameters, in the manner the present knowledge allows.

As the corresponding sound parameters are not directly measurable by the usual methods and devices, new measuring concepts and methods are suggested, based on a family of complex signals, and a test system is briefly presented.

1 Introduction

This paper is an analysis of the state of the art and an attempt to prove the need for a new measuring philosophy in high quality broadcast, recorded or synthesised sound, with emphasis on digital sound. Theoretical and experimental work devoted to human acoustical perception is briefly reviewed in order to derive perception-related parameters of natural and artificially (re-) generated sounds. Some proposals for such parameters areargued, and methodological and technical aspects are discussed.

The following Section exposes the preliminaries to the problem, mainly describing the state of the art in measuring and in digital sound. Section 3 discusses the main physiological acoustics phenomena. Section 4 is devoted to some theoretical issue such as time-dependent spectrum and statistical characteristics.

* Polytechnic Institute of Iasi
† Institute of Medicine of Iasi
§ Polytechnic Institute of Bucharest, Romania

Sections 5 and 6 introduce new measuring parameters and emphasise methodology and technical aspects and applications, while Secion 7 concludes.

2. The Digital Sound and the Quality Concern

It is recognized from the earlier phases of the use of the digital sound (ref 1 and 2) that its main advantages are:

i) the capability to provide high quality for both recorded and transmitted signals, and
ii) the possibility to use the pwerful digital technology on various stages of signal handling.

The compatibility with digital technology allows some other advantages: i) the possibility to correct errors using some kind of redundancy, thus enabling the use of poor quality channels for high quality transmissions, or, in the recording domain, the copying without distortion of degraded recordings; ii) the ease of signal compression and reconfiguration, using more or less sophisticated techniques, including artificial intelligence techniques (ref.3); iii) compatibility with existing or developing digital networks; and iv) compatibility with modern sound synthesis methods.

Even if high quality is not the primary concern, voice and entertainment digital services could qualify for 'all-digital' and automatic tests purposes to use means to measure sound quality specific to the digital techniques. Indeed, the recent advent and the future development in the next decade of new telecommunications services, like ISDN,, ISDN - IBCC, Intelligent Networks etc. (e.g. ref.4,5 and 6) could economically justify the shift from analogue techniques to digital ones. Moreover, as it is already argued (e.g. ref.7), digital test methods have advantages even in the field of analogue systems testing.

On the other hand, high quality entertainment services/broadcasting, such as the (digital) QPSK stereo - sound with TV experiments started in 1986 by BBC, and other experiments coming soon, and music on-demand services (ref. 5), obviously require appropriate measuring

parameters and methods to check for subjective quality.

Despite these ecnomical and technical factors justifying a definite step ahead, there is little progress in the field. Probably, as the TV image has been considered from about 30 years ago to represent the progress in consumer broadcasting technology, less attention is paid to sound. However, one forgets that high-quality, natural, inexpensive sound is a goal never reached, and, moreover is not at all well understood.

In fact, in the high quality sound domain, there exist no similar approaches to those produced 50 years ago in the telephony and illustrated by the work of von Bekesy, Zwisloski and others. Even the researches specifically relating to the digital sound quality refer only to voice, as related to telephony channels (e.g.ref 8 and 9), or to synthesized speech, in the rather low-quality range of the techniques. Consequently, there are no specific tests and standards for high quality sound. Part of the subtle characteristics of auditive perception remain unclear. This handicap is not without consequences; however, it is a subsequent product of the lack of farsightedness of the industry in defining and supporting a clearintended, basic research in the field of high quality sounds perception.

Of course, this paper is not intended to fill this too large gap; it just draws the attention and exemplfies a point of view (H.N.T.).

3. Basic Results in Physiological Acoustics

This section review aspects of physiological acoustics to establish the basis for measurement methodology. An approach from the physiological acoustics point of view is jistified as it is well recognized that usual tests fail appropriately to describe high quality sounds, and the final characterizationi remains with subjective tests (e.g.ref. 8 and 10)

The acoustical findings most relevant to our problem are summarized below.

i) There is well-established evidence that the hearing system is strongly nonlinear in all its subsystems.

ii) The response (to a given sound) is strongly dependent on the acoustical context. It must be stressed that the context-dependent response is shaped at the level of all the stages of the hearing system, i.e. it is not, as sometimes believed, a purely psychic effect. E.g., in the middle ear, and adaption to high level sounds is produced by the stapedius and tensor tympani muscles. Their action attenuates the response for low frequencies (in the region of 300 Hz) and generates an apparent gain in the frequency range 1 to 3 kHz (ref.11).

iii) Despite the general belief - well accomocating low-quality sound transmission - that phase is an unimportant term in sound perception, this is not true. For example, it is proved that "phase effects in narrow band stationary signals at low sensation levels generally occur if there are more than two frequency components in the signal". They are perceived "as changes in the timbre and variations in the prominence of the 'basic pitch'" (ref. 12). (See below for the definition of the basic pitch).

iv) There are great variances in the subjective responses, related to various parameters of the signal. For example, two of the most classical rules, that of loudness and that of duration-intensity intergration, exhibit such subject-related and sound-related variances (e.g. ref. 13 and 14.).

Some more details are useful.

- The kinds of nonlinearity at low levels (near the threshold) differ from those at high levels. There are various direct and indirect proofs of this fact, for example annoyance reactions from noise exposure and noise-induced temporaly hearing loss (e.g. ref. 15 and 16).

- The inner ear specific acoustic, context-dependent response is known as the 'masking' group of phenomena, including 'charactersitic frequency' related changes. Basically, masking consits in the lack of perception of a sound of level over the 'normal' auditive threshold, if the sound is presented after another, stronger one.

! Nonlinearities give rise to 'ghost-sound' perceptions, (also subject to masking-type phenomena!), much influencing the quality of the sound as perceived. The best-known phantom sound is the 'missing fundamental' (or 'basic/residue pitch'). In general, harmonic tones give rise to ghost perception of frequencies related to original frequency differences, i.e.combination tones of frequencies (n+1) f_1 - nf_2 etc.

- Mechanical inertia and thresholds in the subsystems of the hearing system contribute to produce 'instante spectra' in a manner difficult to model and not very well understood. On this subject, references are few and rather confusing. An approach is detailed in the following section.

- The temporal and the frequency discriminatory power of the hearing system are both frequency and context dependent.

All these peculiarities must be reflected in testing principles if perception-like results are to be obtained.

4. Spectral Analysis and Related Problems

Spectral analysis is usually performed in connection with linear systems, in order to derive their transfer functions. However, the spectre of the input andoutput signals are

powerful signal representations, even if the related linear theory is not applicable. Furthermore, this representation fits well enough the frequency discrimination function of the ear.

However, the question naturally arising is: what must be understood by frquency analysis, taking into account the following two basic facts: i) natural sounds can be considered neither stationary, nor ergodic, as their finite and individualized character is too strong;
ii) the hearing system works as a finite time spectrum analyser. It follows that the use of the classical Fourier transform is meaningless due to the infinite duration integration which is involved, and other definitions are to be considered.

The first possibility is to use the so-called 'instantaneous spectrum', introduced by a derivative; however, this is not a measurable value, and what is more important, no connection exists with the hearing organ. (Remark: Attention must be paid not to confuse the above derivative of the spectrum with respect to time with the so-called 'sharpness' of the sound. This 'subjective' property is generally defined as the product of the gradients in the time frquency domain of the sound.)

A measurable and well suited definition for digital signal processing was introduced in ref. 17, and the reader is refered to this.

Finally, in order to suit the nonlinear frequency analysis and discrimination function of the ear, we have introduced an ad hoc definition for the discrete case, involving nonlinear factors. With these empirical factors equalling unity, this corresponds to the spectrum as determined by a bank of band pass filters. The properties of this 'momentary' spectrum will be described elsewhere.

5. Basical Measurement Methodology

To select the parameters to be measured on the transmitting/recording digital system, it can be profitable to start with a point of view quite different from the classical one. Let us remark that the usual manner of measurement is based on the main assumption that the system is linear, (and it follows that only parameters for a linear system are measured), or the system departure from the linear model is negligible. In the latter case, some unspecific, global nonlinearity parameters, are introduced as distortion coefficients. Linearization has the advantage of simplicity both in theory and measuring procedure. It is no longer justified if the number of new 'nonlinearity parameters' becomes excessive, and their meanings obscure the true nature of the system in hand.

Note that the present state of the art corresponds to a great number of 'distortion coefficients' (see for example ref. 18 and the related bibliography), and none of them suitably describes the subjective-like quality.

When low-quality sounds are transmitted, it is not difficult to admit departures from linearity of 1 - 10%, and to linearize the system. However, when errors less than 0.1% are considered and the error of the mean systems are above this limit, it is more natural and economically justified to deal with the system as a nonlinear one, and to define new parameters to measure the departure from an 'ideal' (acceptable) nonlinear system. This method will be used in what follows.

There are two possible approaches. The first, more theoretically founded, is to use some general non-linear model of the system stet to test the system to obtain its specific parameters, and to derive the parameters of the class of output signals, the input signals being known. There are two drawbacks. The first is that once a model is chosen, no departure from it is allowed for the actual system, as any departure gives rise to uncontrollable errors. The second is the difficulty of translating systems characteristics into the characteristics of the output signals.

The second approach, rather emprical, is to use some class of input signals, to measure the outputs of the actual system and to compare them to the 'desired outputs'. This way was chosen as more appropriate. The input signals were chosen to be the most sensitive to changes, from the point of view of subjective perception of the quality.

6. Practical Aspects.

Some of the proposed signals, and their relation to perception are:

i) quadrature, pure tones; phase spectra (phase measurements on noise) are determined;

ii) rise/decay errors for 0.05 to 1 ms rise time triangular transients (total square error) ;

iii) total output error in the 1 - 3 kHz band, for high level broad-band (50 Hz - 500 Hz) input;

iv) total error 0.1,0.3 and 3ms after a high level broadband impulse (.05 ms decay time);

v) total error in the low-frequency band (up to 500 Hz), the input being a high-frequency (1 - 5 kHz) high-level impulse;

vi) total error in a narrow band, when the input is two pure tones, spaced by 1/16 octave and limiting the narrow band measured;