SOME OBSERVATIONS ON THE RESULTS OF OBJECTIVE AND SUBJECTIVE TECHNICAL REVIEWING PRACTICE IN HIGH FIDELITY.

M.COLLOMS Chartered Engineer ,MIEE ,MAES.
Electroacoustic Consultant and Technical Reviewer, London.

INTRODUCTION

This paper covers some of the practical results obtained in the course of the evaluation of a wide spectrum of consumer high fidelity equipment for the publications Hi Fi For Pleasure, Hi Fi News, Hi Fi Choice and Stereophile. The change in reviewing practice in the UK over the last 20 years is outlined, with the emphasis moving from uncontroversial technical assessments to assessments containing an increasingly subjective content. Where subjective judgments are involved, there is often a call for proof of assertions concerning sound quality. This is not easily obtained and in the case of consumer publications the budget is not available. Is there a solution?

REFERENCES

[l] MARTIN COLLOMS, some 2,500 technical reviews published by Practical Hi Fi, Hi Fi Answers, Audio [UK], Hi Fi For Pleasure, Hi Fi Choice, Hi Fi News, The Absolute Sound, Stereophile.

[2] MARTIN COLLOMS, 'Working in the Front Line', *Stereophile*, pp.5-23, [January 1991]

[3] R. HARLEY, 'The Role of Critical Listening in Evaluating Audio Equipment Quality', *AES Reprint, N.Y Convention,* [1991]

[4] S.P.LIPSHITZ, 'The Great Debate - Some Reflections Ten Years Later', *AES Conference, The Sound of Audio,* [May 1990.]

[5] BOB STUART, 'Understanding Noise and Distortion - A New Approach', *AES Lecture, London,* [September 24th 1991]

SUBJECTIVE ASSESSMENT : CHANGES IN TECHNICAL REVIEWING

Twenty years ago, reviewers were mainly engineers - such personalities as John Borwick, Ralph West, John Gilbert, Gordon J. King and Angus McKenzie were all well known. Several had produced handbooks on servicing, audio theory and test technique. They were all technically qualified and highly experienced people who cared about measurement as a science. Continuous refinement and increasing standards would surely track down the details of electrical performance which they believed accountable for any remaining audible flaws. These were held to be responsible for any minor differences in sound quality between various models of audio equipment, except in the case of loudspeakers. The audio world appeared to be well ordered . Distortions were approaching vanishing point, while matters of frequency and amplitude could be qualified to a high degree of accuracy and a panoply of impressive vital statistics could be produced.

Then as now, an audio oscillator, a test load and a volt meter were sufficient to describe many amplifier parameters. These include damping factor, output impedance and continuous output power under a variety of conditions - for example with channels driven together or singly, into various loads and at different frequencies. Power bandwidth may be determined together with overall gain and input sensitivity plus the respective impedance and channel balance. Crosstalk between inputs can be investigated as well as the thermal performance of the power amplifier stage. If the reviewer feels sufficiently brave then the protection circuitry can be checked; for example, by applying a dead short to the output terminals under full power.

For many designs this is a case of being prepared for a fast retreat. Set to dc, the voltmeter can assess the presence of unwanted dc offset at any output. Comprehensive checks of frequency response can be made, including specified equalisations such as the RIAA standard for analogue disc replay. If the luxury of a moderately versatile oscillator is available and could be switched from a sine to square wave output, a power amplifier can then be assessed for overall and reactive load stability, by examining the characteristics of overshoot and ringing on the pulse responses obtained.

All of this is achievable with just a few hundred pounds of test gear. Add a modest distortion analyser and you might think that you could also be a technical reviewer. Defining the above basic parameters is certainly of value in weeding out the poorly designed and audibly inaccurate equipment. What is to be done if much of the product available today returns a high technical performance judged by these standards, but nonetheless exhibits significant audible differences?

Modern audio instrumentation is certainly costly. Establishing a test laboratory capable of covering a wide range of product will cost upwards of £50,000 if good control and hard copy facilities are included. With devices such as wide dynamic range multi-tone generators and distortion meters, high resolution FFT based spectrum analysers and the like, a new and fascinating range of data can be generated, further refining the art of audio product analysis. It all looks very impressive on paper, but only rarely has it provided a significant insight into the more subtle subjective quality variations observed.

The fundamental problem appears to lie in the use of deterministic signals for measurement. We know where they started and where they are going to, and so does the amplifier and presumably its designer. It seems extraordinary that deterministic signals should define the performance; in practice audio equipment is almost never called upon to handle such signals. Speech and music is non - deterministic. Neither past history nor future can be predicted from an examination of a given measurement, except in the crudest sense, for example by looking at a script or a musical score.

APPARENT CONTRADICTIONS BETWEEN CLASSICAL ENGINEERING THEORY AND SUBJECTIVE RESULTS

Engineers raised on an electrical theory based purely on determinism want to believe that their audio world is safe and secure, and that all they need to do is follow the rule book. Electrical engineers who embark on a career as a designer in the audio field often receive a rude awakening. They quickly discover that while the classical theory, if skilfully applied, will produce an amplifier which measures very well, it does not necessarily mean that the end product will pass the subjective tests. In this context, the standard is established by the competition. In the discerning sector of the audio market, sound quality is what counts, and while it is difficult to measure, there is no question about its value and its important role in the commercial success or failure of a design.

Recently there was a fine illustration of this phenomenon which also served to clearly demonstrate the sensitivity of several of the better reviewers to sound quality differences. The case did not relate to some exotic and highly expensive amplifier - the unit in question was a modestly priced integrated amp selling at £200. The manufacturer had a reputation for achieving above average sound quality, although not necessarily with the very best engineering standards. The design was exhaustively researched for a market sector which is known to be highly competitive in the UK. At the end of a process known as subjective tuning , the circuit had received its final adjustment with the available parts and a batch of 50 identical amps were made.

HI FI REVIEWING : OBJECTIVE AND SUBJECTIVE

The company was justly proud of their creation which looked set to lead the pack. Reviewers and dealers alike were widely sampled and a strong concensus as to its remarkable sound quality quickly developed. Then tragedy struck. The manufacturer had planned from the outset to have most of the parts sourced from the Far East - printed circuits, capacitors, resistors etc. All of these had been defined by most careful negotiation, and backed by measurement, to meet or exceed the performance of those parts used for the prototype and pre-production run. But a complete amplifier using the production parts had never been made or auditioned. In the event, the prototype and production amplifiers measured the same, but they sounded quite different. Huge effort was put into trying to find out what lay behind the difference but with no immediate result. The company was already committed to production and high volume orders were arriving with every post. The amplifiers had to be dispatched. To a man, the experienced subjective reviewers knew immediately that what was supplied for official review publication was not as good as the original loan samples. However the cornered manufacturer fought for his commercial position and the ensuing row alienated the press for some considerable time - in fact 18 months. As it happened. while the production model lacked the performance edge of the original, it was perfectly satisfactory in practice and sold quite well, given good marketing and advertising support.

The final stage in this story concerns the admission of fault by the manufacturer, namely his failure to subjectively test the production build. A further 13 months of development, this time using the correct production components, finally brought success. The high quality of those original samples was at last achieved in production, a fact now acknowledged by the subjective press, who showed admirable objectivity despite the acrimonious dispute that arose over the earlier reviews.

Another example concerns a private test where a small panel including the author was employed to differentiate and rank a group of samples of one type of amplifier. All had been previously set up and validated for sound uniformity. Then a series of small changes were made to some of the amps. The listeners had no idea what these changes were, nor did they know which models had been modified. In a disciplined but unstressed listening environment, the amplifiers were differentiated, described and scaled, and the results returned to the manufacturer, who was a much respected major operator in the field. Our results confirmed those they had previously obtained but had not believed.

What level of double blind does this constitute? No one knew what parameter was under test and only after all the work was complete was the nature of the test revealed. It was designed to help choose the best sounding 50 picofarad capacitor type for the feedback loop in the power amplifier. This almost vestigial component value defines the overall phase and stability and is so small that virtually no effect could be sensibly anticipated in the audible range. Yet to this manufacturer's surprise the capacitor type proved subjectively important. All were of similar size, tolerance and rating, and by normal commercial standards, all were of high technical quality. The types included several polypropylene, polycarbonate and polystyrene film components plus an airspaced radio frequency capacitor and finally, a military grade ceramic type. By established review scaling methods for sound quality, these different capacitors accounted for a range of +/-7% in the merit score of the amplifiers, and made the whole exercise worthwhile.

Amplifiers can often be made to sound better by aggressively boosting the engineering budget. However, more often the challenge is to make them better at no extra cost to the consumer. In the case of the above example, the preferred capacitor cost no more than any other and did provide the optimum engineering result. Many tests have shown that individual, well specified passive electronic components can have a direct result on sound quality.

The list is virtually endless and includes volume controls, signal switches, resistors and capacitors - both the small and the larger can type used for power supply reservoirs. Integrated circuits, operational and output socketry , plus printed circuit track and substrates are all influential. No wonder that a wide spectrum of sound quality is heard from the amplifiers available on the world market.

Once the individual passive component is under the microscope, there is no problem in manipulating its physical construction and technology, and consequently its sound. In the case of a wound film capacitor once its basic parameters are understood, then good correlation can be established between sound quality and the absolute quality of the capacitor. Experts in capacitor design know about the deleterious effects of a loose wind containing air voids,as well as the higher resistance of a metallised film or the higher dielectric loss of the poor grade of plastic insulator, never mind the crucial matter of end termination. Thus capacitor sound quality variations are no mystery in themselves. It is all the more surprising therefore, that many classical audio engineers still insist that good amplifiers using such different capacitors must nonetheless sound the same.

LOSS OF AURAL RESOLUTION UNDER STRESSED A/B TESTING

In a recent interesting if somewhat informal experiment, a group of discerning audio listeners were subjected to the experimenter's ideal, a triple blind test. In this case, they did not even know that they were being tested, so there was no stress component in the listening procedure. The listeners were choosing audio components and systems, while unknown to them, the experimenter made changes to absolute phase. While there is solid evidence that the absolute phase of an audio signal is audible, most designers and audio critics agree it is not a major factor affecting sound quality. The point about the test results was that an encouragingly high unconscious sensitivity to phase was shown in the 'triple blind' situation, but when the circumstances reverted to the conventional double blind case, with defined A/B style random sequences for the parameter, the listeners' sensitivity was found to be greatly degraded.

From my own experience, it is often very helpful to instruct listeners not to conceptualise a given class of audio component or components under test, but to keep an open mind for general sound quality changes. This is particularly relevant when assessing a complete audio system. Listeners may be assessing variations in the overall sound through specific changes of signal source,control or amplifying components, a speaker or even an exact speaker location. With practised listeners, working with an open and enquiring frame of mind, without pressure and confident in their ability based on past work, good sensitivity can be demonstrated even with subtle differences between various audio cables.

THE FUNDAMENTAL WEAKNESS OF THE A/B SWITCH

An appreciation of music is largely based on the quality of one's emotional response. In a related way, the appreciation of the replay quality of a given choice of music is also based on the emotional response generated by that replay. For example, it is much harder to judge sound quality using music that you actively dislike. Long experience with listening tests has shown that subjects are most acute when not disturbed by switched, short interval A/B or ABX tests. There are several fundamental problems with switched tests, not least the switch itself. It is surprisingly difficult to remove the influence of the switch components, contacts and connectors from the equation Moreover, some equipment alters in sound quality according to whether it shares a common ground with another active component. This is no mystery since much equipment has hum and noise on its ground lines.

The fundamental weakness concerns the meaning of music itself. Good music carries a message. It has a clear beginning, some sort of development and a conclusion. This can apply to a good rock track of just 3 minutes duration or a phrase, movement or whole work in classical music. Non deterministic, it is impossible to cut up musical phrases with arbitrary A/B switching without severely distorting the music itself.

Such distortion confuses the issue and reduces aural sensitivity almost to zero. Examining the problem in greater detail, there are two crucial aspects. A given piece of music tends to sound different for at least the first 5 or 10 repetitions until it has been thoroughly learnt. Each new hearing subjectively reveals more detail, a greater insight into the structure of the composition and further nuances of instrumental playing and tone colour. There is a wealth of subjective data in just a few minutes of a musical excerpt and one continues to absorb more information with each repetition. Short A/B episodes force the subject to ignore the structure and development. Instead all he can rely on is an unemotional, purely analytical response to variations in level or coloration, or perceived frequency response. Normal human contact with the composer through the music is denied. It is tantamount to assessing a great painting by taking a small square and asking the subject to judge shades of colour and texture with a reference palette.

In addition a gross distortion arises at the instant of switching. The subject is instructed to cut into a musical passage at will. That cut places a false transient edge leading into the following musical section, falsifying the effect. Even with popular music no musical bar is ever exactly the same as the previous one,unless the piece is played entirely by pre programmed machines, in which case it is probably not real music anyway. The lack of absolute consistency from bar to bar undermines the validity of switched A/B testing to a point where one might well expect random results for the subject's sensitivity to small subjective differences. The more one tries to focus on small differences, then the smaller the perception of that difference becomes. Excessive concentration has a numbing effect on a wide range of perception.

Single presentation listening tests imply playing a work from the beginning each time, and playing it for long enough for the subject or subjects to undergo an unstressed, representative emotional reaction to the presentation, whether positive or negative. This is the preferred technique.

OVERVIEW

Looking back over an extended programme of technical reviews - perhaps unique in terms of the close association of the best available measurement combined with sound quality assessment - some radical observations may be made. First, the critical standpoint must be defined. By high sound quality we are aiming to satisfy the needs of a critical listener experienced in high fidelity sound reproduction, and familiar with live music played on acoustic rather than electronic instruments. This standard is not particularly relevant for general purpose speech and music reproduction, though I am convinced that improvement at the top level of replay quality will eventually benefit all those who enjoy reproduced sound.

LISTENING METHODS

Practical, almost day by day experience of auditioning indicates that the greatest sensitivity to fine differences and to general sound quality characteristics is obtained in well ordered but informal listening sessions. The participants should be experienced and relaxed, and have little or no direct interest in the outcome. They have seen or heard sufficient equipment to put any prejudice behind them. In fact, it is not that hard to keep an open mind, and experience certainly helps. High fidelity sound is not an exact science, and some mistakes are made by both designers and reviewers. However one discovers that is virtually impossible to predict the performance of a given unit from its cost or general build and specification. Despite the subjective nature and informal character of most review listening tests, the attitude remains one of dedicated scientific enquiry - the need to know. The goal pursued by the dedicated reviewer and the critical purchaser alike is the approach to perfection in reproduced sound - a more immediate, more convincing reproduction of that favourite orchestra in all its glory.

THE USE OF REFERENCES

It is possible to use an agreed vocabulary to describe what is heard and also to some degree, to score or scale sound quality on a long term basis. The use of reference equipment is vital in reminding panellists of the quality standards involved and also to help scoring. In general, subjective reporting is rather vague and poorly scaled, often due to a lack of references and confidence on the part of some authors.

If the consumer has to rely on subjectively derived opinion, the discipline imposed by scaling imparts greater consistency and comparability to the work.

PRODUCT GROUPS

It is helpful to consider the association between results and listening tests for each product category. These are as follows.

a] Vinyl disc playback including cartridges, tonearms, motors, power supplies and tables.
b] Amplifiers.
c] CD players, transports and digital decoders.
d] Loudspeakers, including stands, spikes, cables.

There is insufficient space to cover all these in detail and the discussion will be restricted to power amplifiers and digital replay. Much data has been amassed on the other categories to show that standard measurement cannot fully account for audible differences.

POWER AMPLIFIERS

During amplifier evaluation, considerable and necessary effort is devoted to the input and output matching characteristics, ensuring that units behave as expected, that the output will be free from either current or voltage overload and that frequency response and linearity errors are negligible. Searching lab tests seek out subtleties in performance at different levels and frequencies. Despite all this, well designed and technically accurate power amplifiers may be differentiated and ranked subjectively. These differences do matter to the more discerning purchasers as well as to technical reviewers, who consider it a duty to rank audio equipment, since the opportunities for falsely representing sound quality are legion.

At the engineering level, it is very clear that the classic deterministic view of a well designed amplifier as a black box with power gain is inadequate. Taking an example amplifier, a textbook approach to power supply design might indicate a 200VA transformer, properly balanced in budget terms at say 10% of the parts bill. Measurement verifies the soundness of the design and the result is auditioned and rated. Now consider a misallocation of the budget. Subjective results indicate that generosity in transformer ratings above the design value is often helpful, even if the rated technical performance hardly changes. Now double the VA rating of the transformer, knowing that much copper and iron is going to waste in pure engineering terms. Given that sound quality is ultimately the arbiter in the real world, the new transformer is found to improve the sound by 15%. This is considered a good return since doubling the transformer size adds 50% to its own cost, and only increases the overall parts bill by 5%. There is not space to explore the good technical explanations as to why over engineering the power transformer aids sound quality, but these explanations make good sense and are not mystical.

DIGITAL REPLAY

When digital replay emerged on the consumer market via the medium of CD, the designers then stated that there were no significant audible errors in digital conversion,either storage or playback. By implication, all CD players were going to sound the same - perfect - and this was actually claimed by some proponents of digital audio. Initially product reviewers were concerned as to whether they would be made redundant by the introduction of CD. After a critical learning period, it turned out that CD was not perfect after all, and while very good test results could be obtained for the system, many perceptive listeners heard defects. In many cases, these were sufficient for them to scale CD below vinyl disc playback.

After several hundred assessments of CD players and digital playback systems, it is only now, after some 7 years of evolution and painstaking development, that the best CD playback is capable of satisfying critical listeners used to enjoying the best analogue mastertape or vinyl disc replay.

Few would deny that there is now a welcome diversity of design and sound quality in the digital audio field and yet experience has also shown that almost no correlation can be observed between sound quality and some of the most searching and painstaking classical, deterministic measurement.

MASKING

Research in recent years has been directed towards finding out how much distortion can be added to an audio signal and still remain inaudible. The definition of inaudibility is a moot point, and is ultimately based on contentious, double blind, A/B listening tests with a statistical criterion for an agreed probability of inaudibility. By implication, a few unfortunate souls will be able to hear it.

Returning to the point, the research has revealed many fascinating aspects of the hearing process. These are now being exploited commercially with the object in view of greatly reducing the cost and quality of a sound carrying channel, while still providing high fidelity when judged from a subjective viewpoint. Masking theory is quite effective at predicting what distortions will and will not be audible, even if the controlling algorithms are not fully perfected. It is fascinating to compare the elaborate dynamic models developed for non linear companding systems with the existing simplistic, classical framework for performance errors in audio equipment. As a general rule, it has been accepted that non linear distortion should be controlled such that total harmonic distortion does not exceed 0.1%, -60dB. With such a limit one is safe in assuming inaudibility, provided that this is the single mechanism potentially affecting sound quality. In fact, if certain criteria are applied, for example, that the distortion be of a low harmonic order, predominantly 2nd and 3rd , then up to 1% is relatively benign - as inaudible as it is visually undetectable on an oscilloscope trace. Conversely if some sharp, discontinuity is present, a zero crossing nonlinearity, or a clipping or squaring mechanism, then here the harmonic order is high and all the distortion energy is concentrated several octaves above the primary frequency range. Such distortion differs from the subtle thinning and sharpening of timbre present with distortion of low harmonic order. The high order effects are amusical and can be likened to a buzz, a rasp or a rattle. Such a noise is clearly differentiated particularly on simple tones or the larger woodwind instruments. Aural sensitivity thresholds of around 0.05% are typical for high order distortion.

Returning to masking, it turns out that if it is properly exploited, a music carrying channel may have distortion levels of many percent. This would be totally unacceptable if judged by classical criteria and yet to a high degree of probability, it will be inaudible to the vast majority of listeners.

HI FI REVIEWING : OBJECTIVE AND SUBJECTIVE

There is a fundamental observation to be made at this point. There is a strong likelihood that the established classical framework for audio measurement is not addressing the right parameters. Moreover, the missing parameters are likely to be based on a complex set of interdependent , processing abilities of the ear-brain combination.This must be the reason why high quality electronics and other audio components can sound different and yet measure almost perfectly by the standards of conventional measurement.

Low bit coding technology based on masking theory provides convincing proof that a level of complex distortion which in itself would be considered audible can in fact be virtually inaudible when it conforms to the rules of masking theory. We know that the ear is non linear, not just from the loudness and sensitivity curves such as those by Fletcher Munsun or Robinson and Dadson. While these indicate intrinsic distortion levels of several percent they do not show that the aural transfer function for phase is asymmetric, responding differently for positive or pressure waveforms as compared with negative pressure or rarefaction. This resulting distortion is predominantly even order which thankfully agrees with our observation concerning our poorer sensitivity to even order distortion, 2nd and 4th for example. This non linearity also has the ability to provide information on the absolute phase of a wide frequency response audio transient. A positive pressure wave does sound different to a negative one.

Given that our best interpretation of the hearing mechanism is to consider it as an array of parallel processing filters, there are also masking and overlap properties dependant on time, level and frequency.

From this, it is clear that the classical static or steady state model used for the analysis of audio equipment performance is inadequate. Nonetheless, those basic, classic test parameters may be ascertained with simple equipment, and it would be unfortunate if fine differences in sound quality were to be generated by relatively trivial linear or non linear errors in the performance of a given audio component. Classical lab measurements are still worthwhile in review, to help weed out those first order factors.

THE WAY FORWARD

I consider that the laboratory mill has processed enough equipment which satisfies normal lab testing in full but which then fails to meet a high enough standard under subjective testing. This indicates some fundamental research needs to be undertaken to explain this apparent contradiction. It is no longer enough to say 'do more double blind tests and the differences will go away.' True, in many cases under such conditions the differences disappear or become meaningless, but this is not the point.

Take the case of dedicated consumers of Scotch whisky who have a favourite brand - all whisky may taste broadly similar, but its origins will endow each brew with its own subtle blend of flavours. You could destroy one of life's pleasures by demanding that a number of discerning whisky drinkers subject themselves to blind testing sessions of at least 15 trials apiece, and then statistically prove there was no subjective difference between the samples. If you are really cold hearted, you could then go on to say that these critics and consumers had been hoodwinked into believing in a massive fraud concocted by all participants in the production and distribution of Scotch. In my view, this attitude is as implausible as the view currently held by several senior and influential academics in the audio field. They state quite clearly that well designed amplifiers meeting traditional test criteria and working within their operating limits, are subjectively indistinguishable . They also assert that even careful and conscientious reviewers must be deluding themselves when reporting continued subjective differences between many such amplifiers.

These reviewers are seen as willing or at best unknowing participants in a similar vast conspiracy. This time, the world's specialist hi fi industry is accused of defrauding its customers by charging excessive sums for a level of over engineering which confers no sound quality benefits. Many of these critics are unwilling to entertain the idea that such sound quality differences could exist.

FUTURE MEASUREMENT

In the development of compressed data coding systems the performance of the ear was the ultimate arbiter. By attempting to model the hearing characteristics, some understanding of the complex multi faceted aural characteristic was achieved. By whatever means the data is compressed, squeezed, shaped and processed, provided that it would still fit the combination, the aural lock could be successfully opened. That pattern is vastly different to the concept of simple, isolated steady state measurements of the amplitude or the distortion of a single sine wave. What is required is some computer analysed technique which could compare an input signal of any kind - multiple tones or music - with the output from an audio component, using the hearing characteristics as a measurement filter. Ideally, this measurement could be scaled in percentage impairment and correlation could be developed with the experience of a wide range of equipment.

Much fundamental research is now being done by many organisations in this field, but unfortunately not with the idea of improving reproduced fidelity but rather the aim of saving money on information storage and carrying channels by bit rate compression. When this paper was being written, there was no immediate prospect of a universal definition for the aural model. It is also conceded by most proponents that masking theory is at an early stage of development and is largely incomplete. Without a solid perceptual model, codec theory and the presentation of a measurement method which correlates well with sound quality will remain out of reach.

However one researcher who has taken a great interest in perceptual models recently reported some progress in this area. This concerns an analytical mask which could be applied to any stage in the audio chain provided that the conditions of use are scaled in terms of real world loudness. With loudness defined the model then provides a basis for predicting the audibility of various distortions both simple non linearities and more complex envelope and noise modulation varieties.The term distortion refers to a multitude of evils and should not be associated only with simple harmonic effects.The better perceptual models indicate that there are many types of distortion whose audibility varies greatly according to the accompanying stimulus.

By continuing the practice of careful listening tests and incorporating a measurement analysis based more closely on the actual hearing mechanism it is anticipated that a useful level of correlation may be established between subjective and objective reviewing of good quality audio equipment.