LOUDSPEAKERS AND ROOMS: 50 YEARS OF RESEARCH

Floyd E. Toole Retired. Consultant to Harman International.

1 INTRODUCTION

A review of the history of audio is informative, but also humbling. A few thoughtful people had the right ideas a very long time ago; the one that stands out in my mind is Hugh Brittain[1], who 80 years ago listed "the common imperfections of electroacoustic systems . . . roughly in order of importance". Other than some understandable differences in terminology, he anticipated what has been verified in the decades since then. Harry Olsen[2], another pioneer also had correctly identified the technical priorities some 60 years ago. Other insightful people exist. How, then, is it that decades later there are significant manufacturers, international standards organizations, and prominent audio professionals who seem not to be aware of, or believe, the insights that existed many years ago and the considerable scientific and technical knowledge that has been added since those early days? Why is it that in the "information age" it is almost impossible to find meaningful specifications on loudspeakers? Why is it that the musical and film sound arts are not preserved as they should be – why does the production/reproduction "circle of confusion" still exist?

Anechoic measurements have been adequately accurate and precise for a very long time, but in the past, measurements were not comprehensive, and lacked the post processing needed for more direct interpretation. Subjective evaluations are notoriously unreliable unless they are conducted under blind or double-blind, conditions, and subject to at least basic experimental controls. As a result, unintentionally Ill-informed opinions have provided much of the guidance in the audio industry.

2 STUMBLING INTO A LIFETIME OF AUDIO RESEARCH

2.1 The Origins

I did my doctoral research at Imperial College, University of London, graduating in 1965. The topic was sound localization, inspired by Prof. Colin Cherry who was intrigued by the stereo soundstage. My experiments used headphones to be able to independently control left- and right-ear signals. When subsequently hired as a research scientist at the National Research Council of Canada (NRCC) I planned to continue the experiments using the excellent anechoic chamber there. For these I needed loudspeakers, and when I measured some highly regarded products the results were depressing. Could they sound as different, and as bad, as they looked?

Being a HiFi enthusiast, in early 1966 I conducted a "Friday afternoon experiment" comparing several products. Fortunately, I knew enough to do the comparisons "blind", and equally loud. The results surprised everybody who participated. There was almost perfect agreement about the preferred loudspeaker, and it happened to be the one with the smoothest, flattest measurements on *and* off axis – Figure 1. I still have the hand-written response sheets. Where was the evidence of personal preferences – at the time, and even now, believed to be a justification for differences in loudspeakers?

The best of the group was a loudspeaker I bought in England as the basis for my first post-graduation stereo system: a KEF Concord. However, it clearly had problems. I found the cause – flexure along the vertical axis of the "racetrack" B139 woofer. A phasing plug and a redesigned crossover yielded a much improved loudspeaker: Figure 1(f). In subsequent tests with many other products it remained "king of the hill", a status apparently warranted by its superior measured performance on and off axis.

What was learned? Listeners in blind listening evaluations are attracted to loudspeakers with flat and smooth on axis response that is somewhat maintained off axis. Evidently human listeners pay attention to both direct and reflected sounds – the room is part of the experience.

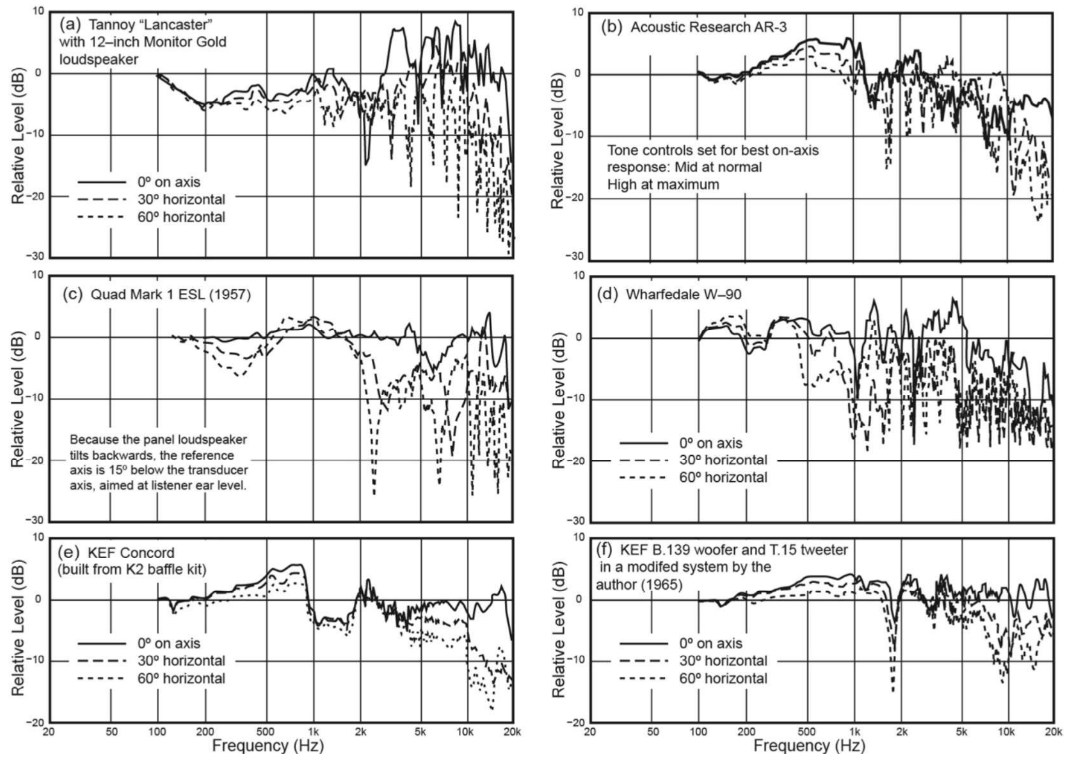

Figure 1. On- and off-axis anechoic measurements on five loudspeakers that were highly respected in the mid 1960s. The sixth (f) was (e) as redesigned by the author. Low frequency data are truncated because the anechoic chamber had not yet been calibrated. This is Figure 18.1 in Toole[3].

To me these observations were consequential. A thorough literature search revealed that there were strong disagreements about which loudspeaker measurements were the most relevant to sound quality. They included:

- on axis (direct sound), which can only apply in acoustically heavily damped spaces like some custom broadcast and recording control rooms.
- A weighted combination of on- and off-axis radiation (direct and early-reflected sounds) This acknowledges a space that is somewhat reflective: e.g. typical domestic rooms, home theaters, many control rooms, and most mastering rooms.
- total radiated sound power (the sum of direct and all reflected sounds). This can only be relevant in highly reflective spaces that generate well diffused sound fields; spaces that are hostile to quality listening.
- a steady-state room curve, in which loudspeaker and room contributions are irrevocably merged and usually octave or 1/3-octave smoothed. The conjecture is that an omnidirectional microphone and spectrum analyzer, that are blind to incident direction and time of arrival, are equivalent to two ears and a brain capable of complex binaural processing.

At that time, these greatly contrasting objectives all had vocal advocates, as discussed in Toole[4].

2.2 The Methodology

It was clear that the definitive psychoacoustic tests had not been done. Those tests had to involve accurate and comprehensive anechoic measurements and well controlled, double-blind subjective evaluations. They also had to involve examinations of the sound fields in typical listening spaces and loudspeaker measurements that can permit estimates of the sounds that arrive at listeners' ears in rooms.

It was critical to this endeavor that I have access to a good anechoic chamber and a budget to construct useful listening rooms. The NRCC made these available to me, and, in my subsequent work at Harman International, I was able to construct even better facilities including a computer controlled positional substitution apparatus to stabilize loudspeaker/room interaction effects.

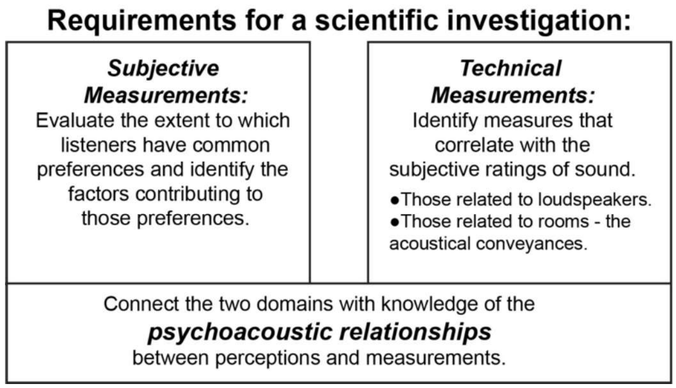

Figure 2. An illustration showing the equal importance of reliable, repeatable, subjective data and technical measurements capable of revealing evidence of sounds that have perceptual importance. Figure 2.3 in Toole[3].

In 1983 I was able to assemble what was to be the key component in the investigation: a computer controlled measurement system that permitted post processing of the anechoic data. The much beloved Brüel and Kjaer analog boxes were inadequate. The new capability included the ability to calibrate the low-frequency performance of the anechoic chamber, using tower and ground plane measurements as references. Useful accuracy was available from 20 Hz to 20 kHz in one sweep.

34 measurements were made on polar and equatorial orbits around loudspeakers: 15° increments in the front hemisphere and 30° in the rear. Spatial averaging of these allowed visual identification of highly audible resonances (not very sensitive to microphone location) while usually innocuous acoustical interference was attenuated by the spatial averaging. One could adjust focus from direct sound (on axis or listening window) to specific angles related to potential first-order (early) room reflections. Sound power calculated from spherical-surface-area-weighted measurements was a final bonus. All of this was done with 1/20-octave resolution. This may have been a first in the industry.

Subjective opinions justified the term “measurements” because with randomized, double-blind, equal loudness comparisons among three or four loudspeakers at a time listeners of all backgrounds were able to deliver “fidelity” ratings with remarkable repeatability. There was impressive agreement among the population. Only listeners with hearing loss departed from the pattern, and they were as distinctive in their opinions as presumably were the defects in their hearing organs. This was discovered when doing evaluations with professional recording engineers for whom hearing loss is an occupational risk[5]. Henceforth, audiometric examinations became a part of listener selection. More recently Sean Olive[6] developed a training program to improve the ability of listeners to recognize and describe resonant colorations, and in the process was able to provide guidance for the selection of program material that is most revealing of these problems. There are some surprises.

The next logical step in trying to understand the perceptual process involved an attempt to predict the sounds that would arrive at a listening location in a room, Figure 3.

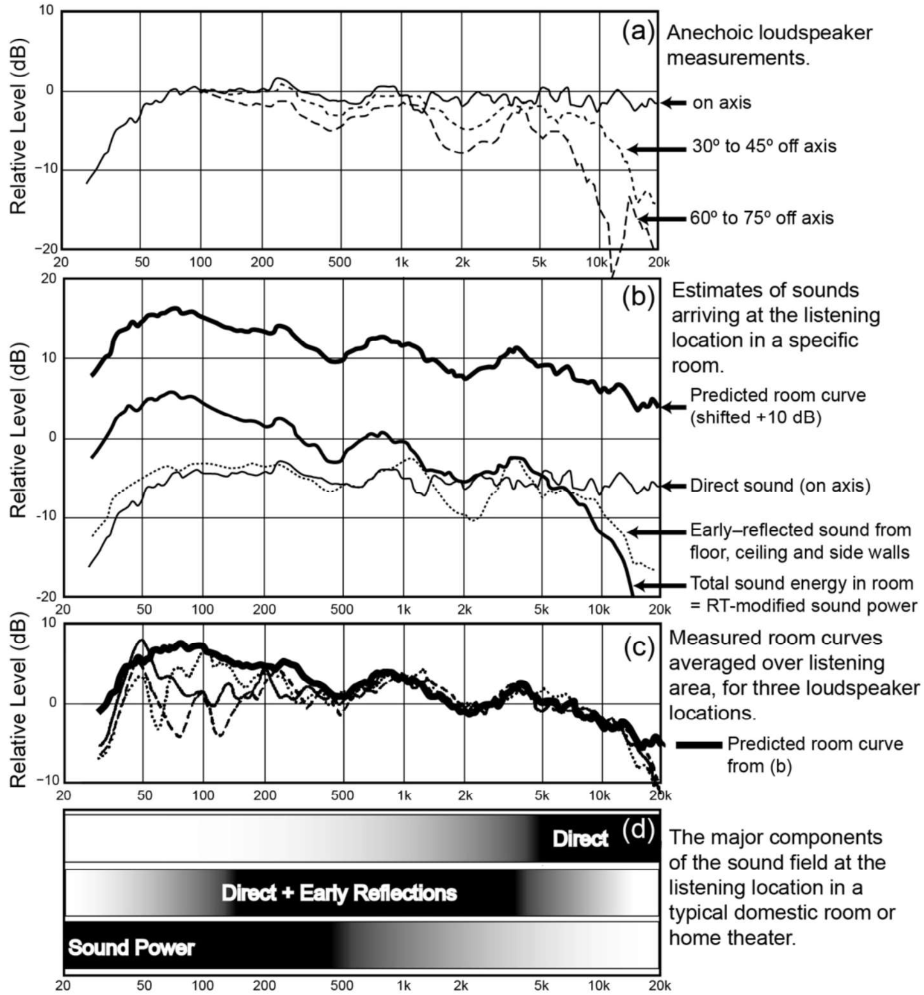

Figure 3. A prediction of a room curve from anechoic measurements and a generalized description of the sound field at the listening location for a KEF 105.2 loudspeaker. Data from Toole[4], Part 2 as adapted for Toole[7] Figure 4. Detailed explanations are there, and in Toole[3], pp. 116-118.

Characterizing a loudspeaker in order to predict what it might sound like in a room involves more than a single curve. In fact, at different frequencies, everything has an effect. This loudspeaker was designed to maximize on-axis performance, and a good job was done, but off-axis performance was allowed to deteriorate. In controlled listening tests the off-axis misbehavior dropped its subjective ratings. Attempting to equalize the room curve to a smoother, flatter shape would destroy the one feature that the engineers aimed to optimize – the direct sound. The solution: it would seem to be a better loudspeaker – one designed to have flattish, smooth on-axis performance and somewhat similar off-axis performance. Subsequent research has confirmed this.

As time passed, many products were evaluated. Figure 4 shows anechoic data on loudspeakers from 1986, organized according to the “fidelity” ratings from double-blind listening tests[4].

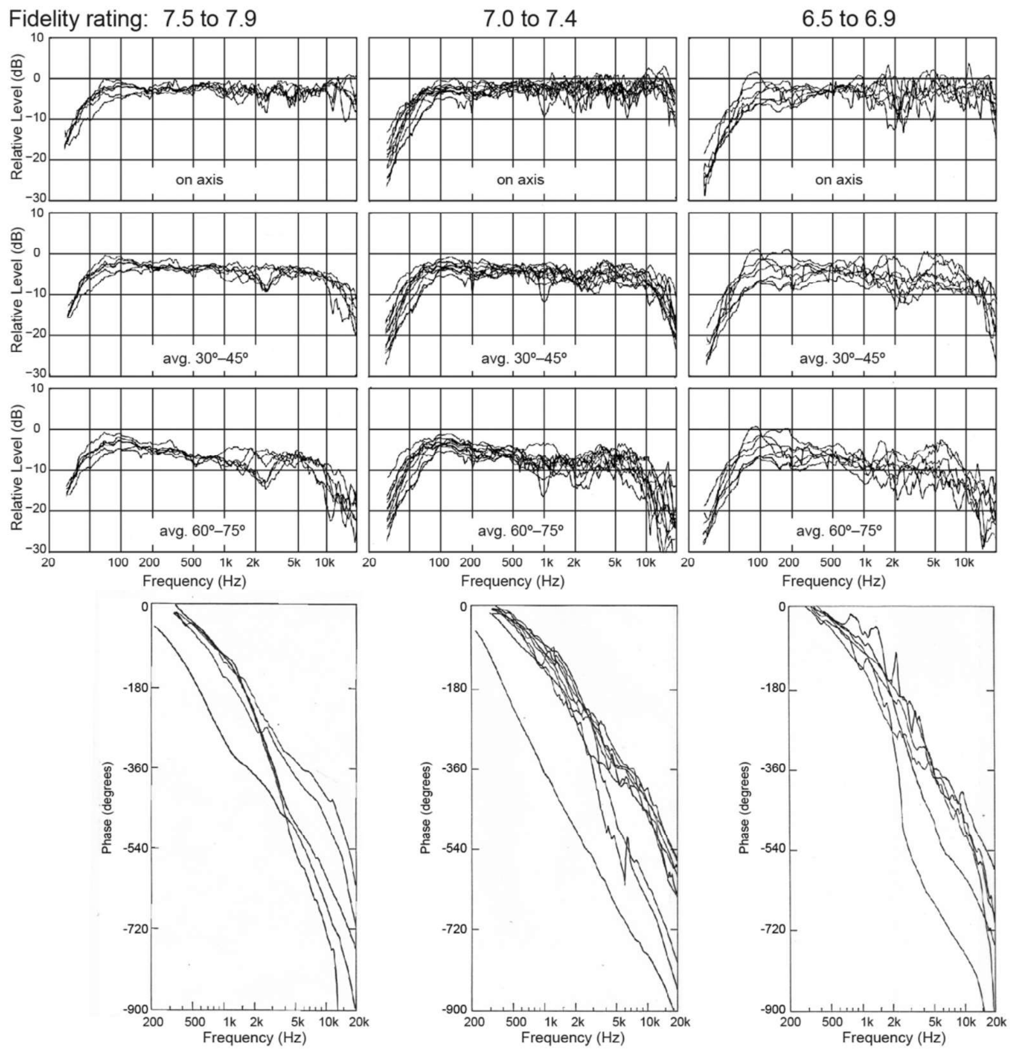

Figure 4. A sample of results from Toole[4], Part 2, Figures 7 and 13, showing loudspeakers grouped according to subjective "fidelity" ratings in three categories. There were six loudspeakers awarded ratings between 7.5 and 7.9, eleven loudspeakers in the range 7.0 to 7.4, and seven loudspeakers in the range 6.5 to 6.9. The original data include a fourth, lower, category that adds nothing to the discussion. The measurements are unsmoothed, 200-point, log-spaced stepped-tone anechoic measurements (1/20-octave). To eliminate the effects of loudspeaker sensitivity the vertical positions of the frequency-response curves were normalized to the mean sound level in the 300-3000 Hz band. This same frequency band was used to normalize listening levels during subjective evaluations. The phase responses may include a small residual time-delay error that could affect the slopes of the curves but not the contours and fine structure. Figure 5.2 in Toole[3].

Some trends are worth noting in Figure 4:

- There is an underlying "flat" trend in these clusters of on-axis frequency-response curves. The variations, even the larger ones, seem to be fluctuations around a horizontal line for the on-axis groups, and around gently sloping lines for the off-axis groups. Crossover imperfections can be seen in several of the curves.
- The low-cutoff frequency progressively decreases as the fidelity rating increases. The listeners liked bass that extended to lower frequencies, not more bass, in the sense that it

was boosted. This has been confirmed in Sean Olive's persuasive subjective/objective correlations[8,9], which showed that bass performance accounted for about 30% of the factor weighting in subjective evaluations.

- The best sounding loudspeakers all had smooth, gently undulating phase responses. Lower ratings were given to loudspeakers with resonances that showed up as discontinuities in the curves; high-Q resonances as sharp discontinuities. It is known that phase response, *per* se, is not an audible factor, but knowledge of phase performance is essential to the design of properly functioning crossovers; the contributions from the low-pass and high-pass portions must acoustically sum smoothly.

Clearly a ±3 dB numerical description does not do justice to the highest rated (7.5 to 7.9) loudspeakers, even in these unsmoothed high-resolution curves. It is evident that smooth and flat on axis was the design objective for all of these loudspeakers which came from different designers, manufacturers and countries. Smooth and flat direct sound was the performance attribute found most attractive to listeners – even though the evaluations were done in a normally reflective room (the prototype for the original IEC 268-13 (1985) room – RT = 0.35 - 0.4 s above 200 Hz).

Deviations from this target are seen at woofer frequencies (below 150 Hz), in the off-axis curves in the woofer/midrange to tweeter crossover region (1 to 5 kHz) and in the tweeter diaphragm breakup region above 10 kHz. Thirty years later these loudspeakers would not be embarrassed if compared to many products in today's marketplace, although today's best designs are definitely better.

2.3 The Present

The NRCC data set was clearly useful, but with faster data acquisition and processing, and automated turntables, it has been improved. It is now what is called the "spinorama" in which the direct, early reflected and later (sound power) information is presented after processing. Directivity indexes are also calculated. This is shown in Figures 5 and 6.

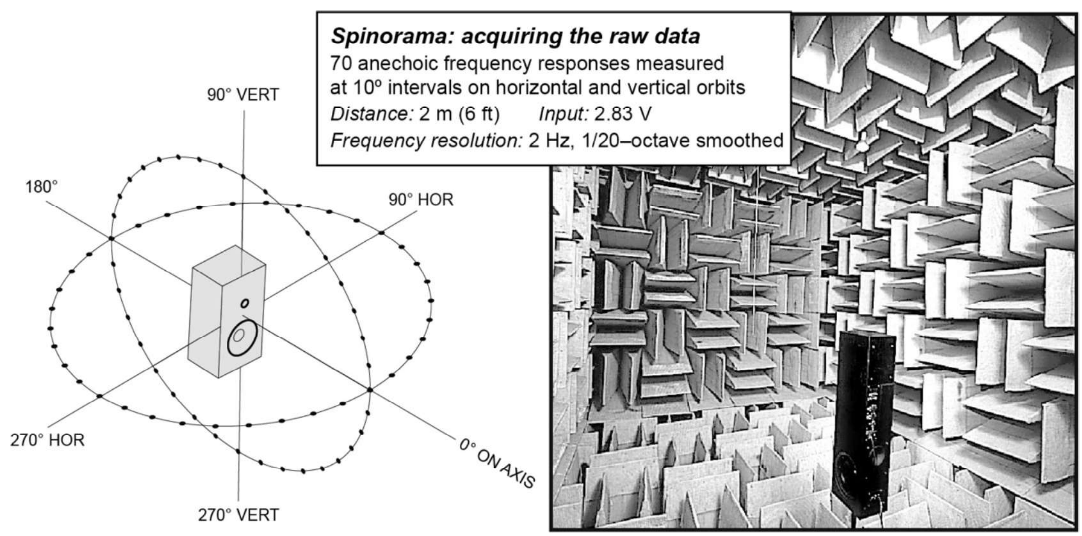

Figure 5. The process used in acquiring the raw data used in the spinorama. The chamber shown has 4 ft (1.2 m) wedges and anechoic ± 0.5 dB at 1/20-octave from 60 Hz to beyond 20 kHz, and has been calibrated to be ±0.5 dB from 20 Hz to 60 Hz at 1/10-octave resolution for the reference microphone and loudspeaker locations. The loudspeaker is automatically rotated at 10° increments for each orbit. The loudspeaker is placed on its side and adjusted in elevation to measure the vertical orbit. Figure 5.5 in Toole[3].

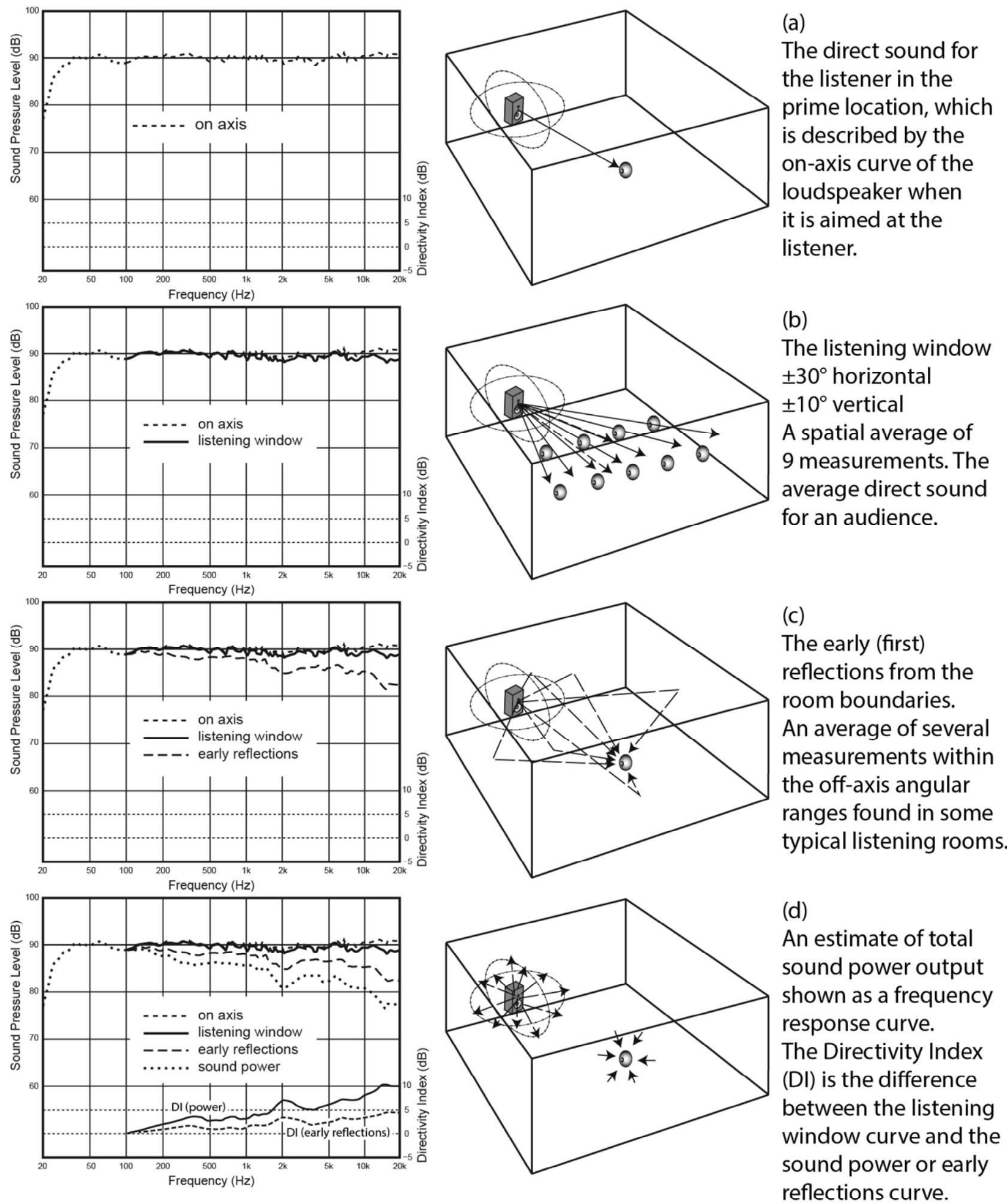

Figure 6. A pictorial description of the individual curves shown in a spinorama. Figure 5.6 in Toole[3].

It is gratifying to see that the spinorama forms the basis for ANSI/CTA-2034-A (2015)[10], that it is being used by at least one prominent non-Harman/Samsung loudspeaker brand (KEF). One audio webzine (audioholics.com) has implemented a do-it-yourself version for some product reviews. It is an optional data output format for Klippel acoustical measurement systems, and other manufacturers have shown interest.

The ultimate test of validity is in correlations between predicted sound quality ratings based on anechoic spinorama data and double-blind subjective ratings of sound quality in a normally-reflective room. This has been convincingly shown by Olive[8,9]. With a correlation coefficient of 0.995 ($p \leq$

0.0001) for 13 bookshelf loudspeakers having similar bass extensions, and 0.86 (p ≤ 0.0001) for 70 loudspeakers of substantially varying bandwidths, sizes and prices. This degree of success indicates that the spinorama data used in the computational algorithm describes the acoustical cues used by listeners in rooms. All of this is the result of years of psychoacoustic research employing the appropriate and accurate measured data, and well controlled, unbiased, subjective evaluations.

Although steady-state room curves, *by themselves*, are not reliable indicators of sound quality, it is significant that these curves can be well estimated from the "early-reflections" curve in the spinorama. Figure 7 shows examples. 1/3-octave smoothing would make them even more similar.

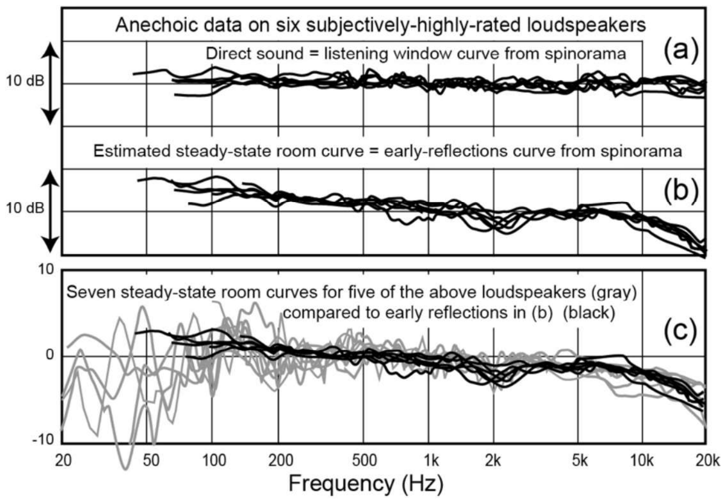

Figure 7. Listening window (a) and early-reflections (b) curves from spinoramas on six highly rated loudspeakers (1/20-octave resolution). (c) shows steady-state room curves for the Revel Salon2, F208, M106 and the JBL Pro M2 measured in five different rooms by four different people using different measuring systems (1/12- to 1/20-octave resolution, four to six microphone locations). The early reflections curves from (b) are superimposed. Excerpted from Toole[3], Figure 12.4.

As shown in Figure 3, the steady-state room curve is dominated by off-axis sound from the loudspeaker. This means that if one sees a steady-state room curve that resembles those in Figure 7(c) the likelihood is that one has chosen a well-designed loudspeaker. If not, equalizing the room curve to have that appearance is *no* assurance that the sound will be comparably good. "Room EQ/calibration" as frequently practiced is flawed, because without comprehensive anechoic data on the loudspeaker one does not know what is responsible for what is measured. EQ may not be an appropriate remedy for a visible wrinkle, or perhaps there is no audible problem. There is a likelihood of degrading the performance of an intrinsically good loudspeaker by equalizing non-minimum-phase phenomena that do not need correction. That said, equalization is very likely essential at frequencies below the transition/Schroeder frequency, especially with multiple subwoofer solutions that can reduce seat-to-seat variations (Toole[3], Chapter 8). Inappropriate equalization is a serious audio industry problem because it has penetrated both the professional and consumer domains.

Placing reliance on steady-state room curves, also called Operational Room Responses (ORR), has a long history, but there was no supporting science. Flat frequency response is accepted as a fundamental target in audio; all electronics have always had it. All loudspeakers that have ever had any merit began with the engineering objective of a flattish on-axis response, which was achieved with varying degrees of success. Because of the frequency-dependent directivity of most loudspeakers the resulting steady-state room curve cannot be flat; it must tilt downward (Toole[7]). However, somehow the notion of flat got misapplied to room curves, and this is seen in the cinema X-curve which is flat to 2 kHz, and in the ORR objectives seen in ITU and EBU documents, which imply a broadband flat response. Toole[3] has much discussion of this topic, including Chapter 11 on cinema sound explaining that evidence that existed prior to the adoption of the X-curve and more accumulated since then all points to listeners in large venues preferring a flat direct sound.

Figure 8 summarizes the situation for small and large listening venues. The data supporting the subjectively preferred film sound curve comes from references 11 thru 14, which show good agreement. The curve from Gedemer[14] represents the combined data well. The curve representing small venue subjective preferences is a smoothed version of the data shown in Figure 7(c). As explained in Toole[7], these curves are what would be expected from typical cinema and home theater loudspeakers radiating flat direct sound into those respective venues. The differences, are partially dictated by loudspeaker directivity and, as indicated, to differences in room reflectivity (reverberation time) at low frequencies and to differences in listening distance (air attenuation) at high frequencies.

In all cases the common factor in subjective preference is a flattish, smooth (i.e. non resonant) direct sound. This has been a consistent feature in decades of double-blind evaluations, seen in Figures 1, 4 and 7. It was a pleasant surprise to find it duplicated in other studies[11,12,13,14] for large venues. Olive[9] and Queen[15] provide further support.

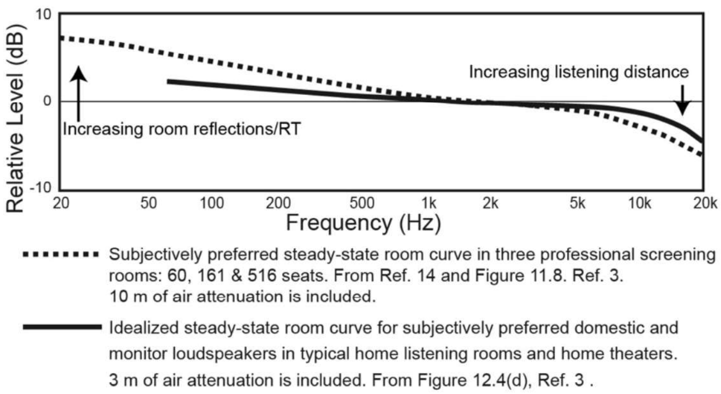

Figure 8. Single steady-state room curves representing collections of similar curves for large venue cinemas and small consumer and professional listening environments. Toole[3], Figure 13.5.

2.4 Conclusion

Direct sound from the loudspeakers should be flat and smooth, which means that steady-state curves in normal rooms must be tilted. Above the transition (small room) or Schroeder (large room) frequency it can be argued that a sufficient criterion for sound quality is that the loudspeaker radiates a flat, smooth direct sound (on-axis or listening window) and a correspondingly smooth, but not necessarily flat, early-reflections curve. Measurements in the room without this data can be misleading, as they can contain irrelevant artefacts. At low frequencies equalization is a necessity.

The spinorama set of anechoic data contains most of the linear distortion data one needs to describe potential loudspeaker sound quality. It is almost certainly more reliable than conventional, sighted – i.e. biased – subjective evaluations in identifying neutral, transparent, reproducers. The circle of confusion, caused by timbre mismatches between professional monitors and consumer loudspeakers, is a significant industry problem. Tone controls are still useful – a “perfect” loudspeaker will not always sound “perfect”. We can do better in both pro and consumer domains.

Equalization of the *anechoic* performance of loudspeakers can result in improved loudspeakers – transducers are minimum-phase devices over their operating ranges. Loudspeakers with dedicated electronics should be the future. However, the features that can benefit from such equalization, mainly crossovers and resonances, are not reliably evident in measurements made in small rooms. In the absence of adequately informative loudspeaker specifications, “room equalization” has become a widely-used fallback. It audaciously claims to deliver impeccable sound from unknown loudspeakers in unknown rooms using a small microphone and analyzer. Obviously, this cannot be equivalent to the elaborate binaural processing of two ears and a brain.

Figures 7 and 8 show room curves resulting from well-designed, subjectively approved, loudspeakers. Some room equalization systems use similar curves as default targets, others don’t, and still others provide optional curves for uninformed users to select from, perpetuating the myth that “we all hear differently”. In any case, a faulty loudspeaker will generate a distinctive room curve, possibly, as seen in Figure 3, due to off-axis misbehavior. Using equalization to reshape the faulty curve to match the “good” one cannot result in the sound quality of a well-engineered loudspeaker. Then begins the exercise of using the convenient user interface to adjust the shape of the room curve target to please the customer. This is using a cumbersome tone-control to make programs of the moment sound subjectively better. Whatever one calls these schemes, it is not a calibration. When standards allow a tolerance of ±3 dB using 1/3-octave resolution, a wide range of distinctive timbral signatures are possible from curves falling within the generous 6 dB (or more) range. It is possible to do much better.

We know how to identify neutral, transparent, accurate loudspeakers in a set of processed anechoically-measured curves. Competent engineers can design such loudspeakers if they are given the tools. When this happens, listeners in double-blind evaluations award them high sound quality ratings averaged over a number of programs. The information is in the public domain and has been for years. Some manufacturers use it, but others appear not to know about it, ignore it, or are not competent to apply it. Evidence of all these alternatives are in products aimed at the professional and consumer markets. It is time to get serious about demanding useful specifications on loudspeakers. The ones available today are, for the most part, an insult to our intelligence.

"The future is already here. It's just not evenly distributed yet." William Gibson

3 REFERENCES

1. F.H. Brittain, "The Appraisement of Loudspeakers", GEC J., Pt.1, vol. 7. pp.266-276 (Nov. 1936); pt. 2, vol. 8, pp. 121-130 (1937 May).
2. H.F. Olson, J. Preston, and E. May, "Recent Developments in Direct-Radiator High-Fidelity Loudspeakers", J. Audio Eng. Soc., vol.2, pp. 219-227 (1954).
3. F.E. Toole, "Sound Reproduction – the Acoustics and Psychoacoustics of Loudspeakers and Rooms", Focal Press, 2017.
4. F.E. Toole, "Loudspeaker measurements and their relationship to listener preferences", J. Audio Eng. Soc., vol. 34, pt.1, pp. 227-235, pt. 2, pp. 323-348. (1986).
5. F.E. Toole, "Subjective measurements of loudspeaker sound quality and listener preferences", J. Audio Eng. Soc., **33**. pp. 2-31 (1985).
6. S.E. Olive, "A Method for Training Listeners and Selecting Program Material for Listening Tests", Audio Eng. Soc. 97th Convention, preprint 3893 (1994).
7. F.E. Toole, "The Measurement and Calibration of Sound Reproducing Systems", J. Audio Eng. Soc., vol. 63, pp.512-541 (2015). This is an open-access paper available to non-members at www.aes.org
8. S.E. Olive, "A multiple regression model for predicting loudspeaker preference using objective measurements: part 1 – listening test results", 116th Convention, Audio Eng. Soc., Preprint 6113 (2004).
9. S.E. Olive, "A multiple regression model for predicting loudspeaker preference using objective measurements: part 2 – development of the model", 117th Convention, Audio Eng. Soc., Preprint 6190 (2004).
10. ANSI/CTA-2034-A (2015). "Standard Method of Measurement for In-Home Loudspeakers", Consumer Technology Association, Technology and Standards Dept., www.CTA.tech.
11. L. Ljungberg, "Standardized Sound Reproduction in Cinemas and Control Rooms", Journal of the SMPTE, vol. 78, pp. 1046-1053 (1969).
12. R.B. Schulein, "In Situ Measurement and Equalization of Sound Reproduction Systems", J. Audio Eng. Soc., vol.23, pp.178-186 (1975).
13. S. Bridges, "Effect of Direct Sound on Perceived Frequency Response of a Sound System", Audio Eng. Soc., 66th Convention, Preprint 1644 (1980).
14. L. Gedemer, "Subjective Listening Tests for Preferred Room Response in Cinemas – Part 2: Preference Test Results", Audio Eng. Soc., 140th Convention (2016).
15. D. Queen, "Relative Importance of the Direct and Reverberant Fields to Spectrum Perception", J. Audio Eng. Soc., vol.21, pp. 119-121 (1973).