Audio Engineering Society

Convention e-Brief 548

Presented at the 147th Convention
2019 October 16–19, New York, USA

Designing listening tests of SR/PA systems, a case study

Eddy B. Brixen[1]

[1] *EBB-consult, 108 Aeblevangen, DK-2765 Smorum, Denmark*

Correspondence should be addressed to Eddy B. Brixen (ebb@ebb-consult.com)

ABSTRACT

It is very common to arrange for comparisons of SR/PA-systems. However, often, these comparisons are organized in a way, leaving procedures less transparent and results rather unclear. Standards for the assessment of loudspeakers do exist. The assessors basically must be trained for the purpose, and the set-up should support double-blind testing. However, in the test of big systems, the listening panel is not necessarily trained, and the practical problems of rigging huge arrays to some degree may weaken the procedures and the results.

This paper describes considerations for the comparative assessment of SR/PA systems. The paper also reports the outcome of an experiment where considered principles were applied.

1 Introduction

It is rather complex to conduct qualified listening tests to assess the qualities of loudspeaker systems offered to venues like clubs, concert halls, theatres, or outdoor festival sites. Most of the time, it ends up as “shoot-outs.” Often these system comparisons or shoot-outs are based on different vendors’ preferences for setup, alignment, selection of audio samples, etc. Many factors are different when comparing to standardized listening tests: the speakers often are visible which may bias the assessors; the audible differences between large systems are greater compared to small-speaker tests; it is easy to place small speakers, but difficult to place large systems; comparisons of systems may even take place with an interval of days compared to the possibility of instant change-over of small systems.

An initiative to organize a shoot-out in the most neutral way possible was taken by Women in Live Music (an international non-profit organization) [1]. In collaboration with Danish Sound Network (at the time a government-supported service for innovation within Danish audio industry) [2] and EBB-consult (a private consulting company), this idea was taken further and developed into procedures, that subsequently were tested (here mentioned as the “SO”).

To the knowledge of the author, there is no independent research published on real-time comparison or quality assessment of PA/SR systems.

Basically, a shoot-out is a listening test on a “bulky scale,” meaning that there are procedures which are difficult to accommodate due to the size of the components, the size of the listening room and the SPL required. This case study describes considerations, the practical solutions, an implementation of a SO-test, some of the results obtained, and the lessons learned for the initiation of research on the subject.

2 Considering the test methodology

There are several standardized methods and recommended practices regarding the assessment of audio [3], [4], [5], [6]. Most of these are related to Broadcast and domestic listening. These listening tests are carried out involving single box loudspeakers which do not take up much space. The listening room and the listening area is relatively small, normally based on recommendations from EBU-, ITU-, or IEC. The SPL during tests is moderate (typically 65-85 dB(A) [6]). When it comes to PA-systems, we may install loudspeaker-arrays with accompanying subwoofer-systems, stacked or suspended, playing at relatively high SPL (typically 75-105) and covering larger listening areas.

One major concern in this kind of testing is how to avoid bias. In general, most PA/SR-engineers attending a listening test have their preferences in advance. It is not very common to apply double-blind testing of PA/SR-systems. However, it was early decided in the workgroup, that the actual test should be carried out as a double-blind test in which neither the assessors nor the test-leader knows which system is playing.

Because each of the PA/SR systems takes up much physical space, it was decided for this SO, that each of the participating systems would be single-array mono systems. Assessing stereo setups perhaps is possible if only two systems are compared. However, the space issue may become problematic if three or more systems are to be compared at the same session.

The venue may have such a geometry that it is not possible to cover the audience area with one single system but rather two or three. By applying only one system (at a time) of course, this single system should only cover a limited part of the audience area. One system of a two-split mono setup should cover half the width of the total listening area; one system in a three-split mono setup should cover at least one-third of the width of the main audience area. All assessors then should be seated inside this covered area.

For the SO it was decided to install the single-array mono systems side by side closest possible to the centerline of the (symmetric) venue.

3 Considering the number of competing systems

To complete a comparison, at least two systems are needed. The major consideration is to find the maximum number of systems to be a part of one test. First, it is a matter of available time. More systems to be compared require more listening time. Also raising the number of different sound examples will extend the time needed for listening. Further, extended listening time creates more fatigue, and thus, less reliable results may be obtained [7]. In practice, if the number of sound systems under test goes up, the number of sound examples must go down.

In the SO it was decided to involve five systems. This decision was pretty much supported by the situation that five vendors had shown their interest in taking part in this shoot-out. Due to the high number of systems, the stimuli were restricted to three sound examples (voice, classical/symphonic music, and rock).

The systems (medium-sized arrays) were suspended from a truss (5.5 m above the stage; stage height: 1.2 m above the floor level) side by side and distributed with a distance to the neighboring system of 1 m center to center. Subwoofers were positioned on the stage below the arrays. It was accepted that the subwoofers of each system could be stacked vertically or arranged as an end-fire solution. It was not accepted to arrange the subwoofers of one system side by side horizontally. The LF-distribution was checked by the vendors, but not documented. However, it was regarded as fair due to the acoustically well-performing room.

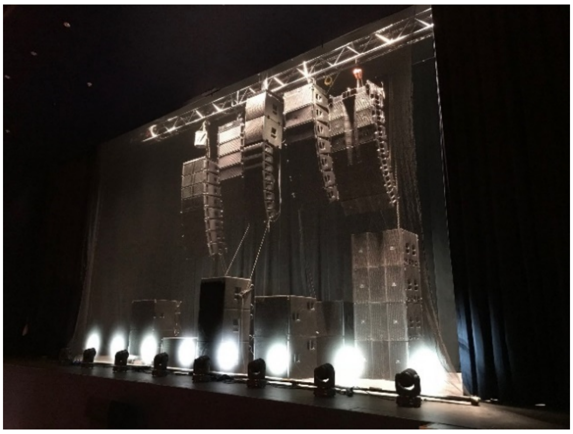

Figure 1. Five systems installed. The main arrays suspended from a truss, subwoofers on the stage.

The final order/positions of the individual system were found by lot-drawing among the vendors.

At the SO the stage-front was covered by Bobinette, a sound-transparent fabric. Front lights and light smoke made it impossible visually to identify any systems.

Assessors were not invited into the room until the systems were disguised.

4 Considering the venue

The purpose of loudspeaker comparisons is mostly related to venues to which it is planned to install new systems. Any venue is special, meaning the optimum setup at one venue may not be the same at another venue.

To ensure the fairest competition, the venue in advance must be well described to the competing vendors. Often, the venue already has information available which is provided to visiting artists and organizers.

Purpose: The description of the venue should include information about the kind of events the venue is intended for, and the system should support. Also, the description should include the priority/percentual distribution of the various events (i.e., rock, classical, theatre, musical, conference, house of worship, etc.). Also outdoor festivals arrange for shoot-outs.

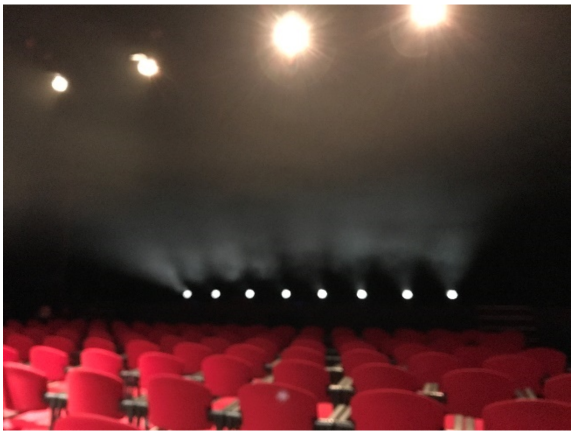

Figure 2. The five systems disguised behind Bobinette, front lights, and light smoke.

Audience size: Information on the audience size, seated or standing, number, areas.

Physical dimensions: This should include dimensions of the room including room volume (with and without stage tower if existing), stage size, rigger points: number and placement, permitted load, trusses, areas of coverage, seating areas, raisers, security areas, possibilities of variable size (i.e., moveable back wall or curtains).

If any 3D model is available, it should be provided to the vendors (i.e., Sketchup, EASE, AutoCAD, etc.).

Acoustics: This information should include reverberation time. If there is a stage-tower that can be separated from the auditorium with an iron curtain, the reverberation time of the auditorium alone and the auditorium + stage tower should be published.

Access: All the usual information on how to access the venue for load-in, phone numbers, e-mail addresses, contact persons, time slots available, etc., should also be given.

Q & A: There should be time for raising clarifying questions in advance to the shootout. Any answer should be distributed to all participating vendors.

The venue chosen for the SO was "Portalen" in the city of Greve, 25 km out of Copenhagen, Denmark. This venue has a volume of approximately 2500 m^3

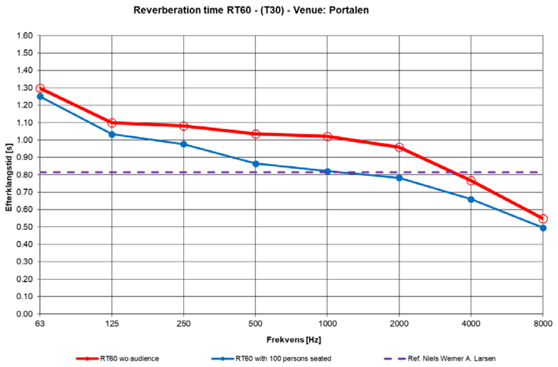

Figure 3. Reverberation time of “Portalen”. Red/bold curve: without an audience but with raiser and chairs. Blue/thin curve: calculated with a seated audience of 100 persons. The dashed line is a reference, according to [8].

and a reverberation time of the un-occupied space, that is is fairly flat around 1.05 s.

Occupied the reverberation time is closer to 0.8 seconds (see figure 3). This is a suitable reverberation time for rock/pop according to research by Niels Werner Adelmann Larsen [8].

The hall is regarded as a multi-purpose venue for most kind of events. The listening area (on a raiser, symmetrical to the centerline) was 16*9.5 m, 11.7 meters from the stage front.

To avoid reflections from an empty floor, rows of chairs were placed in the space between the stage and listening area.

All participating vendors of this SO were provided with information about the venue, including a detailed report about the acoustics (including the diagram in figure 3).

5 Considering the configuration and the tuning of the system

There are several ways to specify PA/SR-systems:

1: The individual components of a system can be specified.

2: The required objective (acoustical) data to be obtained with the complete system can be specified.

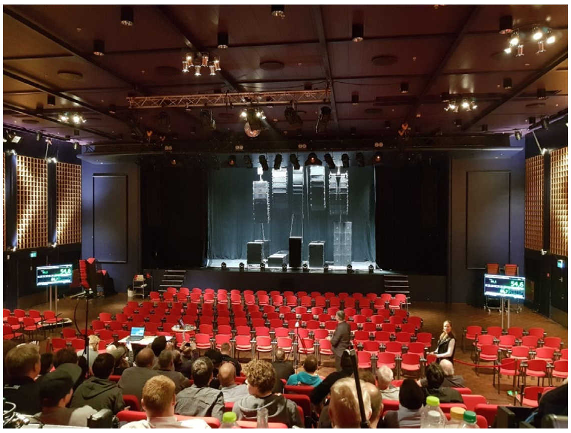

Figure 4. Portalen, the assessors in the predetermined listening area - this picture is taken at that moment when the systems were revealed after the listening tests.

3: A system can be defined and specified by the purpose to which it applies.

4: A combination of the above mentioned.

Defining a system solely by specifying the components does not leave space for alternative and perhaps better solutions. Further, different vendors may not produce identical/comparable components. So, when looking for the best solution, it is a good idea merely to focus on the objective performance data and less on component specifications. However, sometimes, the available space may set some practical limits regarding the size of the components.

Among engineers and users, there are different opinions on how the tuning of a system should be carried out. Especially regarding the tuning and balancing of the low-frequency range. Some want to leave a system with extra bass to make it sound more convincing. So, in a shoot-out, it can be a good idea - before starting - to convince all vendors to perform a system design that meets the needs of the venue and to establish an agreement on the principles of the tuning.

For the SO, the vendors’ task was to provide - and tune - a system that would fulfill the needs of the actual venue. Further, the tuning should be done according to how vendors would leave the system after finished installation. However, one limitation was the positioning of the subwoofers.

The required SPL may vary with the purpose of the system. However, for a listening test, the SPL should be realistic without exceeding any ear-damaging limits. The max SPL tests should not be a part of the listening test. This can be measured separately (C- or Z-weighted with Pink noise or M-noise), for instance, according to NT ACOU 108 [9]. It is meaningless, making listening tests while all assessors are wearing earmuffs.

For this SO, the tree types of stimuli were applied (see section 8) and played at individual SPLs, speech: 75 dB, classical symphonic music: 92 dB, and rock: 95 dB re 20 µPa respectively. These levels were measured as $L_{eq\text{-}A}$ in the middle of the listening area. The systems were adjusted to obtain the same loudness (subjective assessment) around the A-weighted target SPL. A Lo-cut at 36 Hz (24 dB/oct) was introduced in all systems. No other target curves were provided.

Along with the tuning and level setting, some objective acoustic measures can be made for reference and documentation. It is not a good idea to present any measurement before a listening test, as this will cause unnecessary bias.

For the SO, each vendor had a specified time slot for tuning. Further, a system engineer was available setting up a common measurement system (10Eazy) with a display visible to all, for the monitoring of the (A-weighted) SPL (see figure 4).

6 Considering the assessors

The assessors attending listening tests may range in different categories [10]. On one side, we have expert listeners, who are trained, they are familiar with all the attributes which are used to describe the sound. They are in general able to hear and categorize even small impairments in the sound presented.

On the other side, we find the untrained or the naïve listener. This person may have experience of how things should sound, however, is not trained in understanding and scaling the attributes applied.

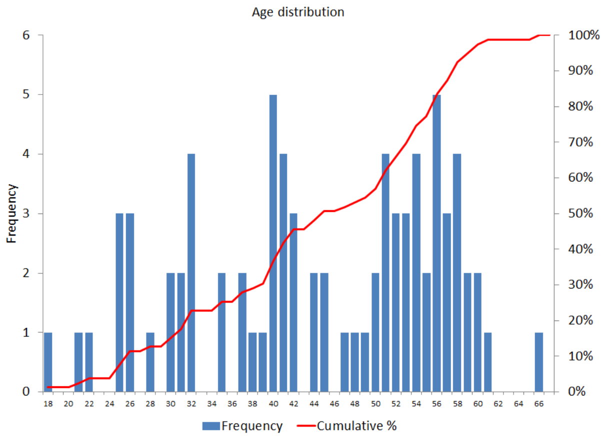

Figure 5. Age of subjects attending the SO.

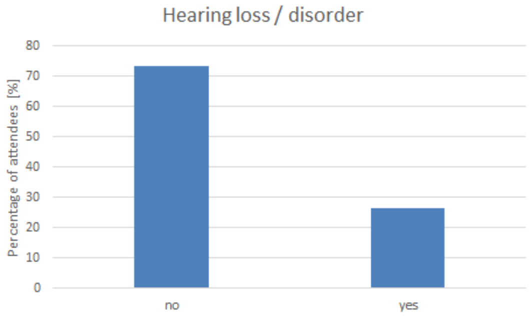

Figure 6. Subjects with acknowledged hearing loss or hearing disorder (approximately 1/4!).

The only thing you can ask an untrained lister is whether he or she likes/dislikes the sound. In the case of AB-comparisons: Which one do you like the best, A or B?

When finding subjects for shootouts, it is often the employees of the venue and affiliated engineers that are "at hand". However, also consultants and "friends of the house" find their way to these shoot-outs, some just because of curiosity. For this reason, it is difficult to ask more than: Which one do you like the best?

In the SO the assessors were invited publicly by announcements on Facebook and in mailings among the members of the organizing organizations.

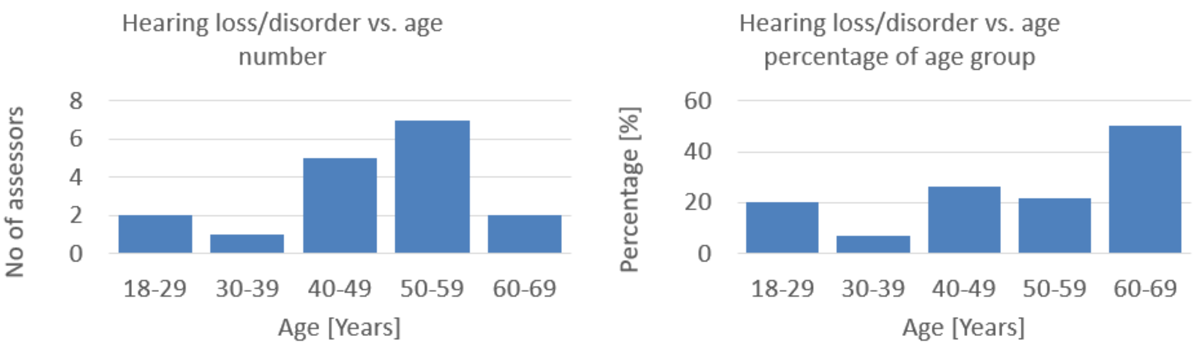

Figure 7. Hearing loss or hearing disorders in dependence of age.

Approximately 80 people signed up for the event — age ranging from 18 to 66 years, median 44 years.

The subjects did not get their hearing tested in advance. However, they were asked whether they were aware of any hearing problems. Also, questions regarding age and experience (years in audio) were included in the questionnaire.

The assessors' age is shown in figure 5, and the response to questions on the hearing disorder in figure 6.

21 out of 79 participants reported some degree of hearing disorder. In figure 7, the distribution of persons with hearing disorder across age is shown.

7 Considering the setup for presenting stimuli

The playback of stimuli, as in this case pre-recorded sound samples, must be very controlled to ensure a reliable result of a test. Also, the randomization is very important in a double-blind set-up. There are different ways to do it, and various companies offer software and hardware for practical solutions. In this SO, the AB-comparison was chosen.

Randomized AB-combinations were executed by Qlab into a digital console (Yamaha CL5), leaving through five auxiliary sends back to the stage box.

The final part of the randomization was a last-minute shuffle deciding which system would get allocated to which channel-number. Only one person - not taking part in the test - would know this combination.

By stepping from scene to scene, it was ensured that all systems 1-5 would be compared against all others two times, so both the AB and the BA situations were presented.

The soundtracks played, were fed into two channels for the AB-test. Channel 1: A, channel 2: B. By doing this, the channels for the A and B systems were defined at all times, reducing the risk of channel confusion.

Coffee breaks were put in between the three sessions to reduce listening fatigue.

8 Considering the stimuli

When presenting stimuli, it should illustrate the properties of the program material they present.

Also, the duration of each stimulus should have a duration that the listener would be able to recognize the type of sound.

Next the listener should have enough time to perform the analysis by listening.

Three types of stimuli were prepared for the SO: Speech, classical symphonic music, and rock.

In the SO, each AB comparison was repeated, providing this sequence: A-B-A-B.

No.	Type	Source	Dur.
1	Speech	Music for Archimedes / Recorded in an anechoic room Microphone: B&K 4003 CD B&O101	16 sec
2	Classical	New York Philharmonic / Avery Fisher Hall Carl Nielsen, 3rd symphony, 1st movement Producer and engineer: Preben Iwan & Mikkel Nymand DACAPO, SACD	21 sec
3	Rock	Artist: AC/DC Album: Stiff Upper Lip Track: All Screwed Up	30 sec

Table 1. Stimuli for the SO. The duration includes the announcement of A or B.

Before each A and each B, the A or B was announced. This was done by audio because it would be difficult to follow any visual cues during the session.

9 Considering the questionnaire

Working with untrained listeners basically leaves us with just one question: Do you like/dislike?

In the SO, it was decided to go a little further by finding some attributes that would have common understanding. The attributes were taken from the sound wheel developed by Torben Holm Pedersen and Nick Zacharov [11]: timbral balance, transparency, and dynamics.

Before the listening test, the assessors were introduced to definitions of these attributes. Also, the assessors were introduced to the stimuli that would be presented to them during the test. During the listening session, the assessors would for each comparison get these questions:

1 Which system exhibits the best timbral balance?
A□ - B□ - A=B□

2 Which system is the most dynamic?
A□ - B□ - A=B□

3 Which system is the most transparent?
A□ - B□ - A=B□

4 Which system do you like the best?
A□ - B□ - A=B□

10 Considering data acquisition

It is not difficult to collect data. A printed questionnaire can do the job. But it leaves the possibility of assessors going back to earlier answers and change them.

The most important thing is data processing. Collecting the data from around 80 people, each answering four questions 40*3 times + additional data on age, hearing, listening seat, etc. ends up with too much manual work. Hence it is important to perform the data acquisition electronically.

For the SO, a questionnaire was created using an Internet-based service: Survey Monkey [12]. The assessors used their smartphones for voting. They could connect using a QR-code or by entering the URL provided for this event. It was checked in advance that the bandwidth of the available wi-fi was enough for trouble-free voting.

For a few without a working smartphone, printed questionnaires were prepared in advance.

The acquired data was available just after ending each session. The results appeared as “1” or “0”: the “1” for the choice (A or B) and “0” for the not chosen (A or B). For the decision “A=B”, the value 0.5 was given to both A and B.

11 Some results

When analyzing the data that comes from simple decisions (one or zero) in a test, there are (at least) two ways of data assessment possible: Either by finding a “winner” of each comparison or by averaging all data. In the first case you may lose information (the winner takes it all). In the second case the result merely mirrors the opinion of the assessors. Thus, the averaged data was applied to this SO.

Figure 8 and figure 9 show the difference between the two methods.

It can be problematic to carry out a listening test if a larger part of the assessors suffers from various forms of hearing disorders. However, analyzing the SO-data did not show any significant difference between assessments carried out by assessors with

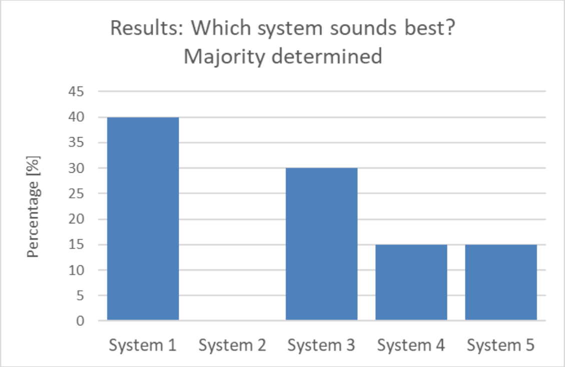

Figure 8. Result, assessment of rock. For each comparison, only the system obtaining the majority of votes got points. In this case the method completely excludes one system.

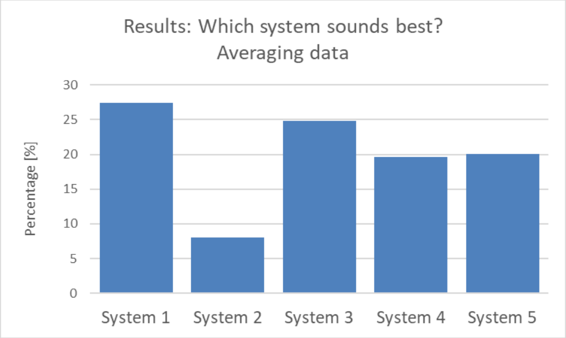

Figure 9. Result, assessment of rock. For each comparison, the A and B system got points based on their share of the voting. This result is very different from the result shown in figure 8.

normal hearing compared to the assessors with hearing problems.

Figure 10 shows the response to the question: “Which system sound best?” presenting data from all assessors and data from assessors without hearing problems only. The difference is so small that it was decided to use data from all assessors.

One system performed less well compared to the four other systems.

That can be problematic in most tests as it may affect the scaling in the assessment of the other systems.

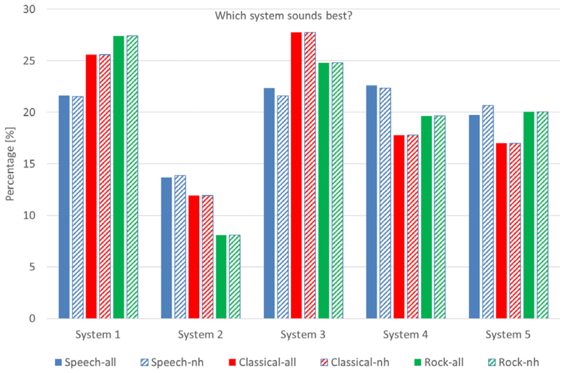

Figure 10. Comparison of data from all assessors (79 subjects “all”) vs. only assessors with (self-declared) normal hearing (58 subjects “nh” normal hearing). The results of the two groups are almost identical.

System 1 has the highest score for rock. System 3 has the highest score for classical music. System 4 has the highest score for speech; even though the results are almost identical to the results of two other systems.

There is a good correlation between the assessment of the three attributes and the final decision: Which system sounds the best. In table 2 The results are listed.

12 Discussion

When designing a setup for a shoot-out involving larger PA/SR-systems, one thing is not possible to overcome: The optimum positioning of the individual system. Even when disguised, after some time during a test, the systems’ positions become obvious although no-one knows the type or the brand of the systems.

It can be problematic to obtain an even LF-distribution from subwoofers with restricted possibilities of placement.

It can be discussed whether binaural recordings can be a tool for the assessment even though the live feeling may get lost.

In this case, each of the vendors had the final decision on how to tune his system.

Signal	Attribute	System 1	System 2	System 3	System 4	System 5
Speech	timbral balance	20.6	13.3	**22.9**	**22.8**	20.4
	transparency	**25.4**	11.5	23.9	20.3	18.9
	dynamics	22.8	15.5	20.4	20.8	20.4
	sounding best	21.6	13.7	22.3	**22.6**	19.7
Classical	timbral balance	25.9	12.3	**27.1**	17.8	16.9
	transparency	26.7	11.7	**26.9**	18.3	16.4
	dynamics	24.2	14.6	**25.7**	16.8	18.7
	sounding best	25.6	11.9	**27.7**	17.8	17.0
Rock	timbral balance	**26.1**	8.4	25.9	19.6	20.0
	transparency	**28.9**	7.6	26.0	18.4	19.0
	dynamics	**27.3**	10.5	23.3	18.9	19.9
	sounding best	**27.4**	8.1	24.8	19.7	20.0
three attributes (average)		**25.3**	11.7	24.7	19.3	19.0
sounding best		**24.9**	11.2	25.0	20.0	18.9

Table 2. Over all results. The highest values are indicated by bold numbers. The bottom two lines show the averaged results of the three assessed attributes compared to the basic question: Which system sounds best?

It can be discussed whether it is better to have a system engineer, that tries to tune all systems to identical responses. Basically, this may not be possible if the systems are based on different principles.

One system did not perform very well. It is always problematic to calibrate scales if one system is very different from the majority.

The shootout showed that the five systems are too many for one testing period. Listening fatigue occurs with a length of the individual sessions of up to 40 minutes.

Presenting only three types of stimuli seem to be too few when testing versatile systems.

Conclusions

A double-blind listening test involving five PA-systems and 79 assessors was designed and carried out.

The assessors were basically naïve listeners. They were introduced to system attributes before the test.

Several assessors indicated having hearing problems. However, their assessments were not excluded from the data.

Five mid-size array systems with subwoofers were tested.

The alignment and tuning were carried out by each of the vendors' engineers. The systems were disguised behind light smoke and light and placed behind the acoustically transparent fabric.

Three types of stimuli were applied. AB testing was setup with the preprogrammed matrix. The routing was randomized last minute before test-start to secure the blind test. The stimuli (A or B) were announced before playing each sample.

The data acquisition involved assessors smart-phones and an internet-based service.

The conclusion is that five systems are too many for a proper comparison as this only gave time for three types of stimuli (speech, classical music, and rock).

More work must be done in the field, especially on how to overcome the practical problems of simultaneously to obtain an optimum positioning of several loudspeaker systems to provide identical listening conditions.

Acknowledgment

In memory of Jan Larsen, Director of Danish Sound Network 1965-2018.

References

[1] https://womeninlivemusic.eu/

[2] https://danishsound.org/

[3] ITU-R BS.1534-3. Method for the subjective assessment of intermediate quality level of audio systems (06/2014). The ITU Radio Communications Assembly.

[4] AES20-1996. AES recommended practice for professional audio — Subjective evaluation of loudspeakers. Audio Engineering Society.

[5] Toole, Floyd E.: Sound Reproduction. The acoustics and psychoacoustics of loudspeakers and rooms. Focal Press, 2009. ISBN 978-0-240-520094

[6] EBU Tech 3278: Listening conditions for the assessment of sound programme material: monophonic and two-channel stereophonic. European Broadcast Union, Geneve, CH, 1998.

[7] R. Schatz, S. Egger, and K. Masuch: The impact of test duration on user fatigue and reliability of subjective quality ratings. J. Audio Eng. Soc. Vol 60, No ½, 2012.

[8] Larsen, Niels Werner Adelmann: Suitable reverberation times for halls for rock and pop music. J. Acoust. Soc. Am. 127, January 2010.

[9] NT ACOU 108: Acoustics: In situ measurements of permanently installed public address systems, NordTest 2000

[10] Bech, Søren, and Nick Zacharov. *Perceptual Audio Evaluation: Theory, Method, and Application*. Chichester, England; Hoboken, NJ: John Wiley & Sons, 2006.

[11] Pedersen, Torben H., and Nick Zacharov. 'The Development of a Sound Wheel for Reproduced Sound'. In Audio Engineering Society Convention 138, 1–13. Warsaw, Poland: Audio Engineering Society, 2015.

[12] https://da.surveymonkey.com/