3rd International Workshop on Perceptual Quality of Systems (PQS 2010)
6-8 September 2010, Bautzen, Germany

Perceived Sound Quality of Small Original and Optimized Loudspeaker Systems

Adrian Bahne

Department of Engineering Sciences, Signals and Systems, Uppsala University, Sweden
adrian.bahne@signal.uu.se

Abstract

The perceived sound quality of small loudspeaker systems with and without digital optimization was empirically evaluated in a listening experiment. Further, the influence on the results of the presentation order of the two versions was investigated, as well as a self-evaluation of potential use for variance reduction. The systems were optimized by means of FIR filters. The two versions of each loudspeaker system were rated in a paired comparison test for music stimuli. For the purpose of analysis a linear gaussian model was applied, resulting in an interval scale revealing interesting information about certainty and discrimination ability of the subjects.

The test investigated whether linear pre-compensation of rather cheap systems results in a significant improvement of the perceived audio quality in a typical listening situation. The results indicated a significant preference for the optimized version and a significant dependency on the presentation order was detected. The self-evaluation was found to be uncorrelated to the test results.

Index Terms: sound quality evaluation, paired comparison test, perceived audio quality, linear gaussian model

1. Introduction

The sound quality of flat televisions, laptops and small loudspeaker systems is an important issue for manufacturers. Optimizing the performance of small loudspeakers used in these devices is a challenging task. Especially the bass performance of these speakers gains a lot of attention, see for example [1, 2].

In this work, the perceived sound quality of small and cheap loudspeaker systems is aimed to be optimized by the means of linear pre-compensation using mixed-phase filters. The focus is hereby not solely on the low frequencies, but on the complete audible spectrum. Whether this kind of optimization is increasing the perceived sound quality of small loudspeaker systems significantly, has been investigated in a paired comparison listening test with graded response applying a linear gaussian model to the test data for analysis. This results in an interval scale with well defined probabilistic meaning of the difference between two points on the scale, allowing for extended interpretation of the results. The characteristic of the scale includes, for example, information about the certainty and discrimination ability of the subjects.

2. Methods

2.1. Loudspeaker optimization

The impulse responses of three small loudspeaker systems, introduced in Table 1, were measured at nine spatial positions around the sweet spot in anechoic conditions.

Table 1: The three small loudspeaker systems under investigation in the present work.

System	Name	Price (approx.)
1	Logitech R-10	20 €
2	JBL on stage 400iD	200 €
3	Fireant FA-004	230 €

A smoothed model was calculated for each loudspeaker channel and the systems were then optimized by means of mixed-phase filters and the use of pre-specified frequency targets [3]. The target spectra were smooth [4] and flat with some regard to significant overall characteristics of the measured spectra. A bass boost considering the speakers frequency limits was applied. The original and optimized frequency responses for the three systems are shown in Figure 1.

2.2. Stimuli

The stimuli used represent music material that people in the target group, here selected as age 18-25, typically listen to nowadays, see Table 2. Stimuli two and four represent hit music from the top lists provided by a compressed streaming audio codec (Ogg Vorbis q5, streaming at approximately 160kb/s). A high-end recording in wav quality was used for stimulus three. To avoid that the subjects decisions are triggered by differences in loudness rather than sound quality, the levels of the original and optimized versions have been normalized using a sound level meter with dBA weighting and pink noise [5].

10.21437/PQS.2010-22

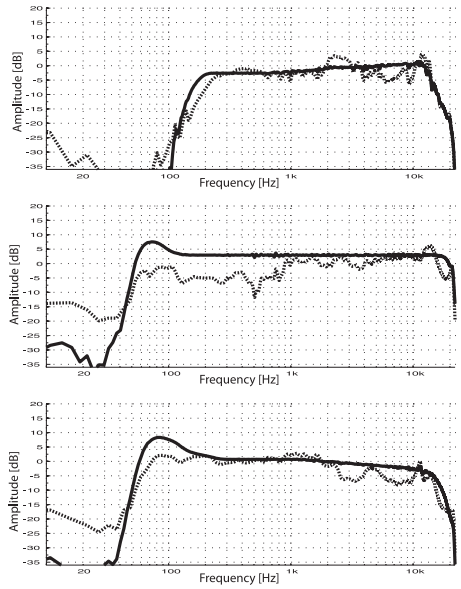

Figure 1: Original (measured, dotted line) and optimized (simulated, solid line) frequency responses of the three small loudspeaker systems under investigation (left channel shown, right channel accordingly). Top: Logitech, middle: JBL, bottom: Fireant.

Table 2: Stimuli used to evaluate the sound quality of the three small loudspeaker systems under investigation.

Stimulus	Artist	Song	Format
1		Bypass (Original)	
2	Milow	Ayo technology	Ogg vorbis
3	Livingston Taylor	Grandma's hands	PCM (wav)
4	Flipsyde	When it was good	Ogg vorbis

2.3. Listening experiment design

The listening experiment was designed as a paired comparison with graded preference response [6]. The test was supported by a graphical user interface, see Figure 2. The subjects task was to rate which of two presented samples, the original and the optimized version, sounds better. Each sample was played for eight seconds, and the possibility to repeat the comparison was given. Each decision was graded in three intervals, labelled *slightly better*, *better*, and *much better*.

Furthermore, the subjects were asked to motivate each rating by means of eleven given psychoacoustic terms: *Bass, mid, treble, natural coloration, tonal balance, clarity, voice, stereo image, distinctness, spatiality, resolution*, see lower part of Figure 2. First, a characterization of the positive properties of the winning stimulus was made. Next, the subjects could select properties that were not better. This was expected to grade the influence

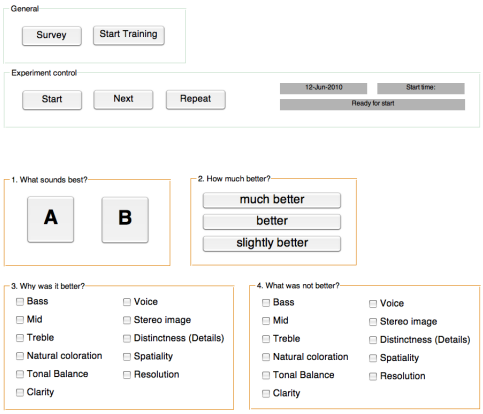

Figure 2: Graphical user interface used for collecting the test data. Two stimuli were played back after each other, and the subjects selected which of them, *A* or *B*, was perceived as overall better and graded their response by selecting *slightly better*, *better* or *much better*. The psychoacoustic terms in the lower part were used to motivate the decision.

of the different psychoacoustic properties on the perception of sound quality for small loudspeaker systems.

The test was *double-blind* and basically a two-interval two-alternative forced choice (2I2AFC) test procedure was used. The test was *balanced*, that is, each sample (optimized or original) was presented equally often in first and second place, and *randomized*. Each stimulus pair was compared either four or six times. If the first four comparisons of a stimulus pair led to the same winner, then the results for this particular pair were considered fully significant and it was not presented any more. For three systems and three music stimuli, the total number of presentations a subject needed to complete the experiment was between 36 and 54 comparisons.

In total, 26 subjects, 12 women and 14 men in the age ranging from 20 to 45, were selected to take part in the test. Neither prior knowledge nor special listening experiences were required to participate. A short training session to get acquainted to the graphical interface and the listening task was carried out with each subject before starting the test. In this session subjects could adjust the overall playback level according to their preference. The playback level was than kept constant. The experiment was carried out in a laboratory akin to a living room.

2.4. Data analysis

For the purpose of analysis a linear normally distributed model was applied to the data, for details the reader is referred to [6, 7]. As shown in Figure 4, the model assumes that a presented stimulus pair (S_i, S_j) excites the sensation $X = (X_1, X_2)$. In the response space for a com-

parison the sensations X_1 and X_2 are assumed to have a variability and are modelled as independent gaussian distributions, both normalized to variance 1 but having potentially different conditional means μ_i and μ_j, given that the stimuli pair (S_i, S_j) was presented. Following [6, 7], the listeners response in a comparison is assumed to be based on the decision variable Y,

$$Y = X_1 - X_2 \sim N(\mu_i - \mu_j, \sqrt{2}) \quad (1)$$

and is quantized to six intervals, depending on the winning sample and the response grading, see Figure 3.

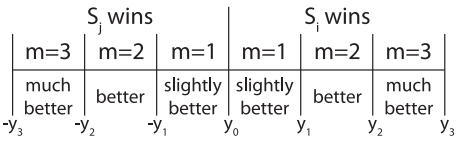

Figure 3: Response quantization. Each response variable Y falls into one of the six response intervals and the response R is quantized according to Equation (2).

$$R = \begin{cases} m & , y_{m-1} < Y \leq y_m \\ -m & , -y_m < Y \leq -y_{m-1} \end{cases} \quad (2)$$

Figure 4: Model assumptions of the linear gaussian model used for data analysis [7].

Defining the quantization limits $y_3 = \infty$ and $y_0 = 0$ leaves y_1 and y_2 for estimation. It is possible to estimate these quantization limits due to the constant variance for all comparisons in the model according to Equation (1). The perceived sound quality of a stimulus S is thus thought of as being the outcome of a normally distributed random variable X with mean μ and variance 1. For a stimulus class $S_{1...n}$, $\mu_1 ... \mu_n$ represent the quality parameters for the different stimuli. The unknown parameters, in the present work μ_1, μ_2, μ_3, μ_4, y_1, and y_2, are estimated from all responses of each subject individually by using a Maximum A-posteriori Probability (MAP) estimation criterion [6]. As $\mu_1 ... \mu_n$ have a probabilistic well-defined meaning, the results can be located on an interval scale.

After conducting the experiment with many subjects and estimating all parameters, the probability that an average listener would prefer a certain sample can be calculated from the distance between two points on the resulting interval scale:

$$P(S_i \text{ is preferred to } S_j) = \Phi((\mu_i - \mu_j)/\sqrt{2}), \quad (3)$$

where $\Phi(\cdot)$ is the cumulative distribution function of the normal distribution. The above described experiment design is, due to its symmetry, fairly *criterion-free* [8] and avoids some of the known response mapping biases, e.g. the *bias due to perceptually non-linear scale* [9].

2.5. Self-evaluation

A self-evaluation with seven questions regarding the subjects listening habits and experiences was conducted. The subjects were asked to rate the following questions on a scale from 1 to 5 or give free answers where applicable:

1. How often do you listen to music?
2. What describes your way of listening to music best? (1: music playing in the background, 5: concentrated listening)
3. How would you rate your hearing expertise? (1: bad, 5: excellent)
4. How important is sound quality for you? (1: not at all, 5: very important)
5. What aspects of sound quality and music are important to you?
6. What kind of music do you listen to?
7. What kind of headphones, amplifier and loudspeaker do you use at home or at work?

The aim of this evaluation was to investigate its possible use for variance reduction of the results, if a correlation between the self-evaluation and the test results could be found.

3. Results

3.1. Overall

The results showed that in general, the optimized version was preferred. An average listener would prefer the optimized version with the probabilities shown in Table 3, obtained by Equation(3). The corresponding values on the interval scale are stated within brackets. Figure 5 shows the interval scales for the three loudspeaker systems. The original is set to the zero level. The percentiles above the original indicate the statistical uncertainties across subjects of the limits y_1 and y_2 of the scale, whereas the horizontal lines show the medians across the 26 subjects for those limits.

Looking at the different quantization intervals for the three scales shown in Figure 5, it can be seen that the subjects used perceptually different and non-linear scales for the different loudspeaker systems. Especially for system two it can be seen that the optimized version is preferred with less significance as compared to the other systems. Furthermore, the scale for system two is more condensed,

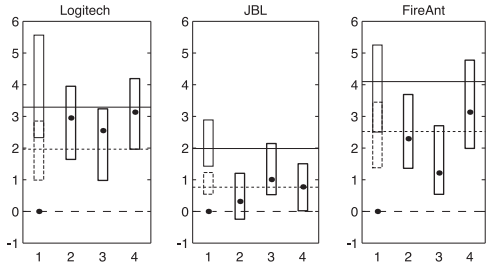

Figure 5: Perceived audio quality of three small original and optimized loudspeaker systems for three music stimuli. X-axis: Stimulus (1: original, 2-4 music stimuli optimized). Dashed lines: Zero level. Fine dotted lines: limit between *slightly better* and *better*. Solid lines: limit between *better* and *much better*. Bullets: Medians across 26 subjects. Boxes: 25th (lower edge) and 75th (upper edge) percentiles for 26 subjects.

which reveals less confidence and discrimination ability. This may suggest that the chosen frequency target for the optimization was not optimal, but the optimization nevertheless improved the loudspeaker system.

Table 3: Probabilities that an average listener prefers the optimized version of the three small loudspeaker systems under investigation. Probabilities obtained by Equation (3). The corresponding value on the interval scale (see Figure 5) is stated within brackets.

System	Stimulus	Probability
1	2	98 % (2.95)
	3	96 % (2.55)
	4	99 % (3.13)
2	2	59 % (0.32)
	3	76 % (1.01)
	4	71 % (0.77)
3	2	95 % (2.29)
	3	80 % (1.21)
	4	99 % (3.13)

3.2. Influence of presentation order

To compare the influence on the results of the order of presentation in the comparisons, the test data was separated into two groups, namely optimized first and original first. Then the linear gaussian model was applied to the two groups and the order-dependent analysis was compared to the overall results.

A significant influence of the presentation order on the results was detected. Presenting the optimized version first resulted in significantly higher preference for the optimized version as compared to presenting the original first, see Figure 6. For the majority of stimuli for all systems the optimized version is perceived as better in general, though with higher ratings if presented first. All systems show a dependency of the presentation order in the scale and results. In particular system two is of interest with the largest differences. For stimuli two and four, the original version was actually preferred when presented first, but with significantly lower ratings, and the overall results indicate the preference for the optimized version.

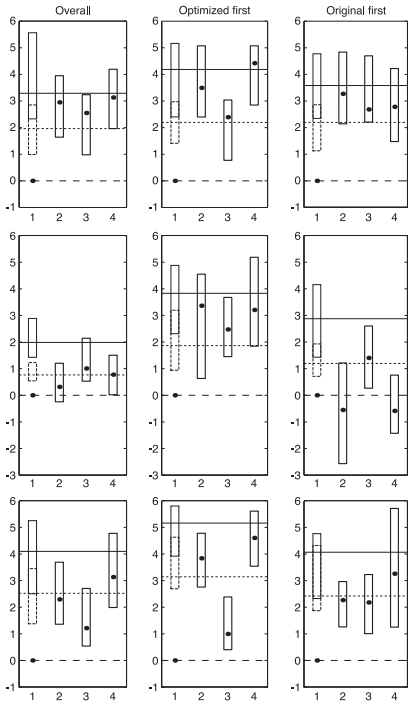

Figure 6: Influence of the presentation order on the results. X-axis: Stimulus (1: original, 2-4 music stimuli optimized). Dashed lines: Zero level. Fine dotted lines: limit between *slightly better* and *better*. Solid lines: limit between *better* and *much better*. Bullets: Medians across 26 subjects. Boxes: 25th (lower edge) and 75th (upper edge) percentiles for 26 subjects. Top: Logitech, middle: JBL, bottom: Fireant.

In combination with the motivations for preference and the measured and optimized frequency responses, the dependency on the presentation order revealed interesting information about the perception of sound quality in this setting. In situations with strong differences in the frequency response, like for system two with very weak bass and mid in the original version, and a short presentation time, most subjects seem to been confused by the discrepancy between the original and optimized version, in particular when the original was presented first. Looking at the motivations for their decisions, most subjects preferred the original because of *treble*, *voice*, *clarity* and *tonal balance*, which were perceived as missing or wrong

in the optimized version. Even this may sound odd, listening to the optimized version first, most subjects indicated the exactly same motivations, this time for the optimized version and with a significantly higher confidence in their decision.

3.3. Self-evaluation and listener performance

The self-evaluation conducted by the subjects before the experiment is found to be useless for variance reduction. No correlation between the answers and the test results could be found. However, the listening test design makes it possible to easily pick *well performing* listeners afterwards. The total number of needed presentations, lying between 36 and 54, is a suitable criterion for this purpose. A low number indicates a listener certain about his/her decisions and a high percentage of congruent decisions. Table 4 shows the distribution of listeners over the possible number of presentations in the current experiment.

Table 4: Possible numbers of total presentations needed to complete the experiment (N_p) and the corresponding numbers of subjects (N_s). The sum of all N_s is 26.

N_s	0	3	4	6	3	2	4	2	1	1
N_p	36	38	40	42	44	46	48	50	52	54

Figure 7 shows the results for the subjects finishing the test in maximum 40 comparisons compared to the rest of the listeners. Besides the higher number of congruent answers, it can be seen that these subjects are also less, if at all, affected by the order of presentation and make a more extensive use of the scale, indicating better discrimination ability.

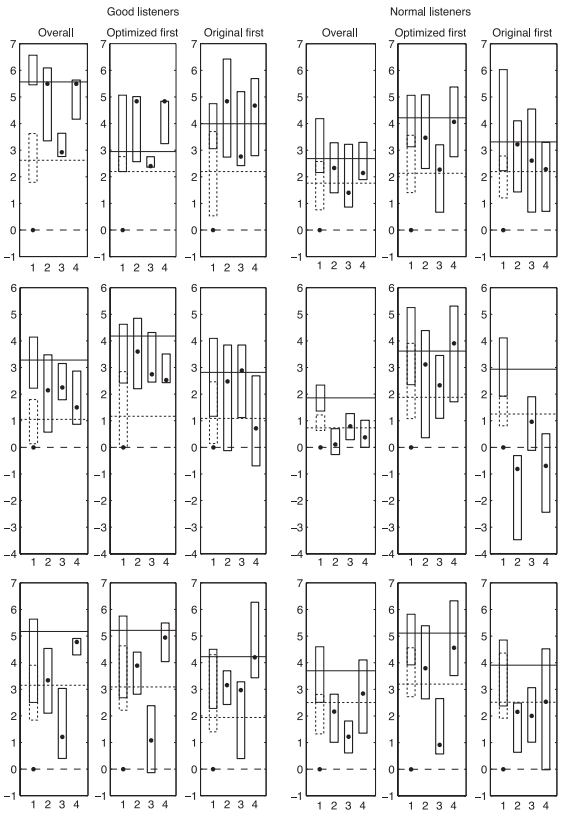

Figure 7: Results for well performing (good) and normal listeners. X-axis: Stimulus (1: original, 2-4 music stimuli optimized). Dashed lines: Zero level. Fine dotted lines: limit between *slightly better* and *better*. Solid lines: limit between *better* and *much better*. Bullets: Medians across 7 (good) and 19 (normal) subjects. Boxes: 25th (lower edge) and 75th (upper edge) percentiles for 7 or 19 subjects respectively. Top: Logitech, middle: JBL, bottom: Fireant.

4. Discussion

4.1. Presentation order

As discussed in Section 3.2, the motivations by means of the introduced psychoacoustic terms (see Section 2.3), revealed interesting information about the perception of audio quality in this setting. System two, for example, showed an approximately 6 dB weaker response in the range of 60-1000 Hz, see Figure 1, middle diagram. Presenting this impaired version first, subjects stated weak performance in *treble* and *mid* for the optimized version, but with low certainty in their rating. Presenting the optimized version first, subjects were sure about the optimized version is performing better in *treble* and *mid*, but this time with significantly higher certainty in the rating. Even though the optimized version was perceived as being beneficial for those two attributes when it was presented first, listening to the impaired version first made it difficult for the subjects to make a decision. The *bass* performance was rated as significantly superior in both cases.

It could be argued that these differences due to presentation order might mainly be due to the short presentation time used. The influence of long-term listening on the results, especially on the effect of the presentation order, is not investigated in this work. This remains an interesting problem for future research.

4.2. Linear gaussian model

As discussed in Section 3.1, the linear normally distributed model revealed that the scale the subjects used was non-linear. To illustrate this effect, an additional analysis applying an ordinal scale [8] with discrete intervals to the test data has been executed. Figure 8 shows both the interval and the ordinal scale for the sound systems under investigation. Looking at the quantization in-

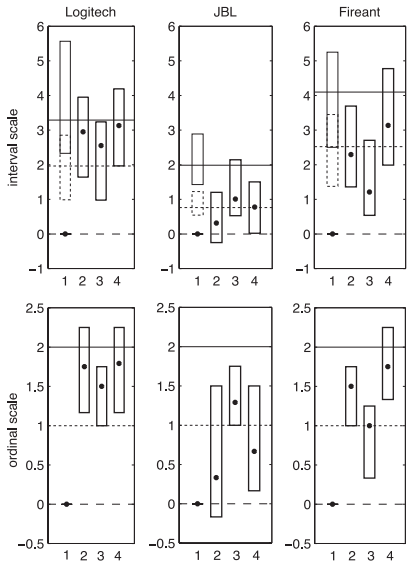

Figure 8: Difference in response mapping between interval and ordinal scale. X-axis: Stimulus (1: original, 2-4 music stimuli optimized). Dashed lines: Zero level. Fine dotted lines: limit between *slightly better* and *better*. Solid lines: limit between *better* and *much better*. Bullets: Medians across 26 subjects. Boxes: 25th (lower edge) and 75th (upper edge) percentiles for 26 subjects.

tervals of the interval scale, it can be seen that the scale is non-linear. Thus, the *bias due to perceptually non-linear scale* [9] has been avoided using the linear gaussian model.

Further the linear gaussian model revealed that the subjects adjusted the scale according to their perception. Figure 8 shows the response mapping by the linear normally distributed model compared to the ordinal scale. Using an ordinal scale, the subjects ratings were mapped to the discrete intervals and all scales are equal. The introduced interval scale indicated that the scale was adjusted by the subjects, and the response mapping is inaccurate using an ordinal scale. In both scales it can be seen that system two is rated lower, but the rather condensed interval scale indicates, in addition, that it was more difficult for the subjects to make a decision.

4.3. Biases

To summarize the benefits of the linear normally distributed model and the applied experiment design, see Table 5 for some of the known biases affecting listening test results [9] and their reduction.

5. Summary

It was shown that the applied linear pre-compensation of small loudspeaker systems increases the perceived sound quality significantly. Further is was shown that the presentation order plays a significant role on how different frequency responses are rated. A set of well performing listeners was not found by means of the self-evaluation but by analyzing the test results. The linear normally distributed model and the resulting interval scale revealed interesting information and was found to be superior to an ordinal scale in many aspects.

Table 5: Some biases reduced by the linear gaussian model and the applied experiment design. For a discussion of the biases and their reduction, see [9].

Bias	Reduction
Recency effect	Short listening time and looped recordings [9]
Bias due to equipment appearance, listener expectations, preference, and emotions	Large number of listeners and double-blind test procedure [9]
Stimulus spacing bias	Indirect scaling given by the paired comparison procedure [9]
Bias due to perceptually non-linear scale	Indirect scaling given by the paired comparison procedure [9]
Interface appearance bias	No direct interaction with the scale which arises from the linear gaussian model afterwards
Range equalizing bias	To some extent by linear gaussian model, still no absolut measure though

6. References

[1] J. Liebetrau, D. Beer, and M. Lubkowitz, "Psychoacoustical bandwidth extension of lower frequencies," in *AES 127th Convention*, (New York, NY, USA), 9-12 October 2009.

[2] P. Minnaar, "Non-linear signal processing for low frequency enhancement," in *AES 34th International Conference*, (Jeju Island, Korea), 28-30 August 2008.

[3] L.-J. Brännmark and A. Ahlen, "Spatially robust audio compensation based on SIMO feedforward control," *IEEE Transactions on Signal Processing*, vol. 57, pp. 1689 –1702, May 2009.

[4] F. E. Toole, *Sound Reproduction*. Focal Press, 2008.

[5] S. Bech and N. Zacharov, *Perceptual Audio Evaluation - Theory, Method and Application*. Wiley, 2006.

[6] M. Dahlquist and A. Leijon, "Paired-comparison rating of sound quality using MAP parameter estimation for data analysis," in *1st ISCA Tutorial and Research Workshop on Auditory Quality of Systems*, (Mont-Cenis, Germany), 2003.

[7] A. Leijon, "Lab 3. paired-comparison rating of sounds," 24 September 2007. Lab instructions, KTH-Speech, Music and Hearing.

[8] A. Leijon, *Sound Reproduction, Introduction and Exercise Problems*. Stockholm, Sweden: KTH Electrical Engineering, 2009.

[9] S. Zieliński, F. Rumsey, and S. Bech, "On some biases encountered in modern audio quality listening tests-a review," *J. Audio Eng. Soc.*, vol. 56, pp. 427 – 451, June 2008.