Evaluation of Designs for Loudness-Matching Experiments

Esben Skovenborg[1,2] and Søren H. Nielsen[2]

[1] University of Aarhus, Dept. of Computer Science, Åbogade 34, DK-8200 Århus, Denmark
esben@skovenborg.dk

[2] TC Electronic A/S, Research Department, Sindalsvej 34, DK-8240 Risskov, Denmark
shn@tcelectronic.com

In a study of the assessment of loudness of music and speech, several different experimental designs were compared. The effect of using a fixed-reference design, compared to using a balanced pair-matching design with various degrees of redundancy and balance, was investigated with a technique inspired by resampling statistics. In a loudness matching experiment using a fixed-reference design, the subjects match all stimuli against a single sound segment selected in advance. Alternatively, in a balanced pair-matching design, both stimuli of a pair are drawn from the same segment collection. In balanced pair-matching, each of the segments do not need to be matched with every other segment; the specification of the set of segment pairs to include in the experiment corresponds to the degree of redundancy and balance in the design. The results indicated that the choice of specific fixed reference segments did influence the accuracy of the assessments – in particular a 1 kHz tone was inferior. When using characteristic samples of speech material, rock/pop, and classical music, as fixed reference, no significant difference in accuracy between the genres was indicated. Applying the balanced pair-matching method with only a small degree of redundancy yielded considerably more accurate results than the designs without repetitions evaluated.

1. INTRODUCTION

The perceived loudness of music and speech can be measured by means of a controlled listening experiment. The loudness of homogeneous sound segments, with a duration of several seconds, may be compared because the overall loudness of each segment is perceived to be fairly constant. The property of the sound which is assessed is its *long-term loudness*. The results from a loudness assessment experiment could be studied to gain knowledge of the perception of loudness (psychoacoustics). The results could also be utilized in the development of objective measures or algorithmic models of loudness.

Measuring the perceived loudness in a listening experiment implicates several choices regarding the experimental design and procedure. In particular, a **loudness matching experiment** could either be based on a *fixed-reference* method, or on what we shall call the *balanced pair-matching* method.

When using the fixed-reference method, a single sound segment is selected in advance as the reference stimulus. The subjects then match all (other) stimuli against this reference. In contrast, when using the *balanced pair-matching* method, no sound segment is distinguished as the reference; both sound segments in a pair of stimuli are drawn from the same collection. The composition of the set of pairs to be matched is said to be *balanced*, basically because the frequency of occurrence of the different segments is the same.

In [1] we described an investigation of loudness assessment of music and speech. A pilot experiment was carried out using half the number of subjects and a scaled-down

collection of sound segments. The reduced number of different segments allowed the experimental design to include a match of every segment against every other segment. In the pilot experiment, the consequence of using different classes of sound segments as the fixed reference was investigated in terms of accuracy of the responses. Furthermore, the effect of using the fixed-reference method, compared to using the balanced pair-matching method with various degrees of redundancy, was investigated.

Numerous (other) previous studies, concerned with loudness matching, have applied the fixed-reference method. As the reference segment these experiments used typically either pure tones, e.g. [2], or noise-segments [3, 4]. In an ongoing study of loudness measurement methods, within the ITU-R, a speech segment was used as fixed reference [5, 6, 7, 8].

A special kind of loudness matching experiments is concerned with *calibration*; for instance, the MEDUSA study of multi-channel level alignment [9, 10], or the work by Aarts in which the objective was – via the loudness equalization – to eliminate loudness differences as a factor in subsequent tests of subjective quality of loudspeakers [11, 12]. Both music and various kinds of noise have been used as stimuli. However, in this calibration type of experiment, typically the same stimulus is used as both the reference (A) and the test stimulus (B).

In this paper, we examine the effect of using the fixed-reference method with various sounds as reference, and compare that to using the balanced pair-matching method with various degrees of redundancy and balance. Our empirical investigation is based on experimental data from the mentioned pilot experiment, analysed with a technique inspired by resampling statistics. To the best of our knowledge, the fixed-reference method has not yet previously been compared to a balanced pair-matching method, in connection with loudness assessment experiments.

2. THE EXPERIMENT

This section provides a brief summary of a pilot experiment conducted as part of a study of the loudness assessment of music and speech – the details of this study can be found in [1]. The pilot experiment was employed partly to verify the functionality of overall procedure, the setup and software, and partly to investigate into certain details of the experimental design. These investigations are described in sections 3 and 4.

2.1 Subjects, Stimuli, and Procedure

A loudness matching experiment was conducted, using the method of adjustment [2, 13]. The subjects were instructed to adjust the loudness of a comparison stimulus (B) using a volume or gain control until it matched a reference stimulus (A). The relative level of one of the segments in each pair was controlled by the subject using an endless rotary knob [14]. The method of adjustment was chosen because it was fast and intuitive, and therefore suitable for an experiment involving a relatively large number of segment pairs to match. Four subjects participated in the pilot experiment. All subjects were trained in audio engineering and thus very familiar with the task of adjusting levels using a knob.

The stimuli for the pilot consisted of 20 sound segments: 6 representative or characteristic samples were selected from each of 3 classes of sound: rock/pop music, classical music, and speech material. In addition, 2 test sounds were included: pink noise and a 1 kHz pure tone (Table 1). Each sound segment was edited into a homogeneous segment of 10-15 seconds duration. The dynamic range of the experiment was controlled by means of a level normalization using a pseudo-loudness function together with a stochastic spread of the presentation level [1]. This procedure also randomised the direction of the adjustment level, to counteract bias phenomena.

In the pilot experiment, every segment was matched with every other segment in a stimulus pair. Additionally, each of the two test-sound segments was matched with every

segment in both A/B and B/A-order. Thus, the relative loudness of a total of $18 \cdot (18-1)/2 + 2 \cdot (2 \cdot 20 - 1) = 231$ segment pairs was adjusted by each subject. The presentation sequence of the pairs and the A/B-order were randomized to suppress the effect of any factors related to the timeline of the experiment. The total effective response time of the pilot experiment was 3.76 subject hours.

	No. of segments
pop/rock music	6
classical music	6
speech material	6
test sounds: 1 kHz tone, pink noise	2
Total number of sound segments	20

Table 1. Sound segments in the pilot experiment

2.2 Analysis Method

A statistical model of the responses from the experiment was developed. The loudness level of every stimulus can be estimated via the model, based on all level adjustments from a given subject.

The ***DifferenceLevel*** variable describes the relative level adjusted by the subject, in dB. Each *DifferenceLevel* value, based on a response obtained in the listening experiment, is associated with the pair of sound segments that were used as stimuli. Therefore, the set of *DifferenceLevel* values corresponds to the loudness level for every sound segment, the ***SegmentLevel*** variable. The value *DifferenceLevel(i,j)* is the relative level of the stimulus pair consisting of the segment with index *i* (as stimulus A, played at a fixed level) matched with segment *j* (as stimulus B, played at a level adjusted by the subject). For example, *DifferenceLevel*(3,5) = 2.3 dB means that the segment with index 3 was perceived by the subject as being equally loud to segment with index 5, when the latter (the B segment) was presented with a relative gain of +2.3 dB.

Let the value *SegmentLevel(i)* be the estimated loudness level of the sound segment with index *i*. The relationship between *DifferenceLevel* and *SegmentLevel* could then be expressed, as follows:

$$DifferenceLevel(i,j) = SegmentLevel(i) - SegmentLevel(j) \qquad (1)$$

The linearity assumption implied by eq. 1 was shown to hold within the relatively narrow SPL range covered in these experiments, roughly the stimuli were between 60 and 80 dB SPL. All the *DifferenceLevel* values resulting from the experiment can be considered as a set of linear equations of the form expressed by eq. 1 – specifically, we get *nTotalAdjustments* equations with *nSegments* unknowns. The data from the main experiment was modelled using a General Linear Model (GLM). The GLM can be regarded as a combination of a Multiple Linear Regression and an analysis of covariance (ANCOVA) ([15] [16]). The *regression* refers to the process of estimating the optimal *SegmentLevel* parameters, given the *DifferenceLevel* adjustments from the listening experiment.

In [1], statistical models were presented which incorporate bias terms and genre-dependent factors. However, in this paper only the simplest of these models, as described in eq. 1, is used: each observation (*DifferenceLevel*) is explained (only) by the difference between two continuous predictor variables (*SegmentLevel*), corresponding to the two segments in the matched segment pair. The ANCOVA of this model showed that *SegmentLevel* is a highly significant set of factors in predicting the observed *DifferenceLevel*.

When the set of *SegmentLevel* parameters is estimated given the combined responses of all the subjects, the *SegmentLevel* values were said to describe the ***common loudness***. Alternatively, the *SegmentLevel* parameters could be estimated for each individual subject,

i.e., without combining the responses, in which case the estimates describe the *subjective loudness*. In this paper only the ***subjective loudness*** is considered in order to avoid the issues of between-subject disagreement [1].

3. THE BALANCED PAIR-MATCHING METHOD

3.1 The balanced pair-matching and the fixed-reference methods

As mentioned in the Introduction, a loudness matching experiment could either be based on a *fixed-reference* method or on the *balanced pair-matching* method. Comparing the two methods, it is observed that by using balanced pair-matching instead of a fixed-reference scheme –

1. The choice of which sound segment to use as reference becomes a non-issue. The investigation of between-subject disagreement of loudness assessments [1], indicated that the disagreement would tend to be lowest for pairs of segments with similar properties (e.g., genre, spectrum, or dynamics). This suggests that it would be difficult to find a single reference sound segment which would consistently yield a low between-subject disagreement when compared to a range of different segments.
2. The bias due to the subjective impression of the particular fixed reference segment is avoided, or at least spread out over all segments. Even in a loudness assessment experiment a subject's judgement might be affected by his or her preference, annoyance, or interpretation (semantic content) of the reference sound segment.
3. All of the obtained level adjustments are used in the estimate of the loudness of every segment, via the statistical model (see section 2.2). Alternatively, when using a fixed-reference scheme, only the fraction *1/nSegments* of the adjustments are used for estimating the loudness level of each segment.

Suppose we have a collection of stimuli consisting of *nSegments* sound segments. In a pair-matching ***full experiment design***, every segment is matched against every *other* segment, requiring a total of *nSegments*(nSegments-1)/2* adjustments. At the opposite, a ***minimum experimental design*** requires only *nSegments* adjustments (e.g., each segment *i* could be matched with segment *i+1*). In this way the *redundancy* in the experimental design can be varied between the *minimum* and the *full experiment* designs. Generally, in any experiment, increasing this redundancy (i.e., obtaining more observations or samples) will lead to a better suppression of the experimental error. Note that in a *fixed-reference* experiment, the redundancy would be increased by repeatedly matching the *same* pairs, whereas in a *balanced pair-matching* experiment, the redundancy is obtained by including more of the $nSegments^2$ different segment pairs.

Figure 1 illustrates two balanced pair-matching experiments and a fixed-reference experiment, each involving 6 sound segments. In the figure, each node in a graph represents a specific segment, and a vertex (connecting line) indicates that the corresponding segment pair is included in the experimental design. In Figure 2 the same three experimental designs are represented in matrix form; a dark square in field (*i,j*) indicates that segment *i* is matched with segment *j* in the experimental design. The full pair-matching experiment, where each segment pair is rated once, in either A/B- or B/A-order, requires *nSegments*(nSegments-1)/2* adjustments, or 15 adjustments in the example 1. The example 2 shows a *partial* pair-matching experimental design with 9 adjustments, i.e., with a size in-between the full design

and the minimum design. In the example 3, the fixed-reference experiment, each segment pair is rated once, requiring *nSegments* adjustments[1].

In all loudness-matching experiments a *variance* or uncertainty is associated with each adjustment. The experimental design determines how these variances are distributed. Suppose, in the three examples (Figure 1), that the variance associated with matching segments 1 and 6 would for some reason be relatively high. In the fixed-reference example (ex. 3), the estimate of the *SegmentLevel* for segment 6 would then be relatively uncertain. In contrast, consider the balanced pair-matching design (ex. 2). In this experiment, any two segments are connected by several paths in the graph (Figure 1). Thus, the variance – such as experimental error – is spread out evenly between the different *SegmentLevel* values, and hence minimized overall.

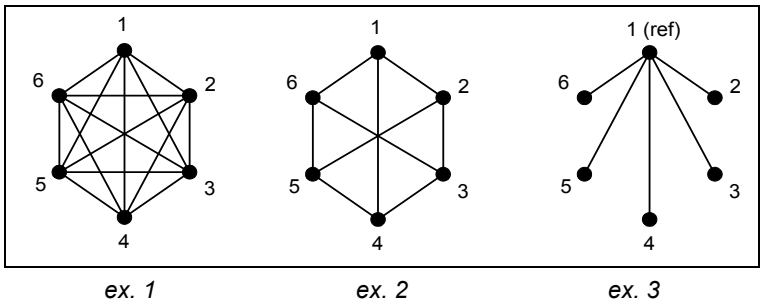

Figure 1. Each of the three graphs illustrates an experiment with a number of adjustments (the vertices) involving pairs of six sound segments (the nodes). *Ex. 1*: balanced pair-matching, full experiment, 15 adjustments. *Ex. 2:* balanced pair-matching, partial experiment, 9 adjustments. *Ex. 3*: a fixed-reference design, 6 adjustments.

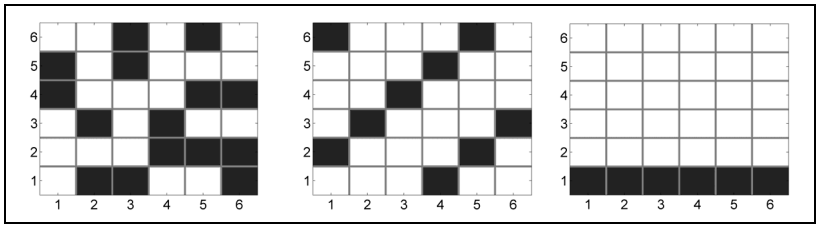

Figure 2. A segment-matrix representation of the three experimental designs as illustrated in Figure 1. In each matrix, a dark square at (*i,j*) means that segment *i* is rated once against segment *j* in the design.

3.2 Requirements of balanced pair-matching experimental designs

The requirements that must be fulfilled for an experimental design to be *a balanced pair-matching* design can be specified formally. Let *D* be a design matrix of the size $n \times n$ in which *D*(*i,j*) is the number of adjustments in an experiment with segment *i* as stimulus A and segment *j* as stimulus B. Let *n* be the total number of sound segments in the experiment (=*nSegments*). For a given design, represented by *D*, the following measures could be calculated:

$$SegmentCountA(i) = \sum_{j=1}^{n} D(i,j) \qquad (2)$$

$$SegmentCountB(j) = \sum_{i=1}^{n} D(i,j) \qquad (3)$$

$$SegmentCountAB(i) = SegmentCountA(i) + SegmentCountB(i) \qquad (4)$$

$$PairCount(i,j) = PairCount(j,i) = D(i,j) + D(j,i) \qquad (5)$$

[1] In some of the previous fixed-reference experiments, the reference was matched with itself, and in others it was not. This detail is, however, probably not of great importance.

To determine whether an experimental design (D) is a *balanced pair-matching* design, the following properties are calculated, based on the above four measures:

$$SegmentCountMaxDiff = \max_{i \in 1..n} \left| SegmentCountA(i) - SegmentCountB(i) \right| \tag{6}$$

$$SegmentCountRange = \max_{i \in 1..n} SegmentCountAB(i) - \min_{i \in 1..n} SegmentCountAB(i) \tag{7}$$

$$PairCountRange = \max_{i \neq j} PairCount(i, j) - \min_{i \neq j} PairCount(i, j) \tag{8}$$

We refer to an experimental design (D) as *balanced pair-matching* when it has the following properties:

$$SegmentCountMaxDiff \leq 1, SegmentCountRange = 0, PairCountRange = 0 \tag{9}$$

However, in our investigations, the above requirements were relaxed to allow balanced pair-matching designs to be constructed for a given number of matches or pairs – eq. 9 cannot be fulfilled for any given n. Henceforth, we shall also designate a design as balanced, when –

$$SegmentCountMaxDiff \leq 2, SegmentCountRange \leq 1, PairCountRange \leq 1 \tag{10}$$

We have developed an algorithm for the loudness assessment project [1] to generate random experimental designs which fulfil the requirements in eq. 10. The main experiment in that project employed this algorithm, using $n = 147$.

In Table 2 the measures in eq. 2 to 8 are calculated for the three experimental designs illustrated in Figure 1 and in Figure 2. The table shows that the example 1 is indeed a perfectly balanced pair-matching design, as it fullfills the requirements in eq. 9. The example 2 is still a balanced pair-matching design, according to the relaxed requirements of eq. 10. The example 3 is a fixed-reference design, and does neither fullfill the *SegmentCountMaxDiff* nor the *SegmentCountRange* requirements for balanced pair-matching designs.

	Balanced pair-matching, full experiment	**Balanced pair-matching, partial experiment**	**Fixed-reference experimental design**
SegmentCountA	2,3	1,2	1
SegmentCountB	2,3	1,2	0,6
SegmentCountAB	5	2,3	1,6
PairCount (i≠j)	1	0,1	0,1
SegmentCountMaxDiff	1	1	5
SegmentCountRange	0	1	5
PairCountRange	0	1	1

Table 2. The different measures to distinguish the balanced pair-matching design, calculated for the three examples in Figure 1 and Figure 2. Where the measure depends on the specific (*i,j*) multiple values are listed.

A given experimental design may be considered as a *graph* as in Figure 1. In addition to the balance requirements stated in eq. 10, we believe that it is desirable for the design if its graph consists entirely of *cycles*, and furthermore that these are of approximately equal length (number of vertices). This property appears to minimize the variance on the *SegmentLevel* parameters. We have not, however, formalized or pursued this idea any further.

4. EMPIRICAL INVESTIGATIONS

This section describes an empirical investigation of various experimental designs. Each experimental design is evaluated in terms of how close its resulting loudness estimates are to

the *best estimate*s of the relative loudness. This way, the fixed-reference method with different kinds of reference segments is examined. Furthermore, the consequence of using the balanced pair-matching method, with various degrees of redundancy is investigated.

4.1 Method of Investigation

In the performed experiment, every subject completed the *full experimental design*, i.e., each subject adjusted the relative loudness of every segment against every other segment. Suppose that the *best estimate* of the loudness level of every segment (the *SegmentLevel*) is the estimate based on the responses from the full experiment, i.e., all adjustments made by a given subject. Using a *subset* of the available responses that corresponds to some smaller experimental design, the deviation from the *best estimate* could then be calculated. *Fixed-reference* experiments could be simulated by sub-sampling of the experiment data. Balanced pair-matching experiments of different sizes could also be simulated. The influence of the specific choice of segment pairs, constituting the simulated experiment, can be considered as a sampling error. To even out this sampling error, the simulation procedure is repeated many times, with different random experimental designs of the same specification. This general method, which is illustrated in Figure 3, was inspired by other techniques of statistical resampling, such as the bootstrap procedure [17, 18].

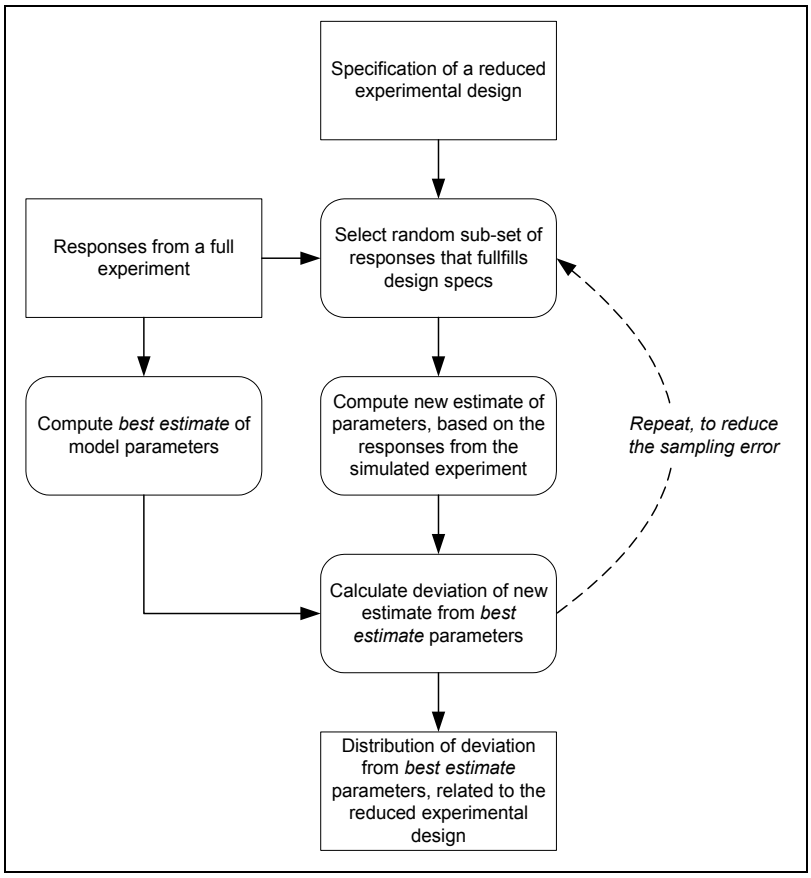

Figure 3. The algorithm that was developed to investigate the effect of various experimental designs on the accuracy of the results, by means of simulated experiments.

Note that the deviation from the *best estimate* is calculated from each individual subject's adjustments only; hence, the between-subject variability is not taken into account. Moreover, due to the varying degree of 'balance' and redundancy in the sub-sampled (simulated) experiments, a reliable estimate of the bias factors was not possible to obtain in every case. Therefore, correction of the subject's adjustment bias and A/B-order bias [1], was not used in the calculation of the *best estimate*s.

4.2 Test-sounds as Fixed Reference

Two *test sounds* were included as stimuli in the performed experiment: a 1 kHz pure tone and pink noise. All segments were compared directly to both of the two test sounds, in both A/B-orders. Therefore, a subset of obtained adjustments can be extracted, simulating an experiment in which a test sound is used as the fixed reference to which all other sound segments are compared. This way, it can be calculated which of the two test sounds, used as fixed reference, would yield the *SegmentLevel* parameters closest to the *best estimate* of the *SegmentLevel*. The particular algorithm that was developed is specified in Figure 4.

1. **Assign a specific segment as the fixed reference.**

   For each test subject:
2. **Using all adjustments, compute the *best estimate* of the *SegmentLevel* parameters.**
   The *SegmentLevel* parameters are determined relative to each other; a fix-point is used to anchor them onto the phon scale.
3. **Select the subset of adjustments corresponding to all pairs of the fixed reference and another segment.**
   This subset will consist of either *nSegments* or *2*nSegments* adjustments, for a reference segment from the speech/music class or the test sound class, respectively.
4. **Compute a new estimate of the *SegmentLevel* parameters, using only the selected adjustments.**
   Furthermore, a constant offset to all *SegmentLevel* parameters is estimated and used to align the new *SegmentLevel* parameters with the *best estimate SegmentLevel* from step 2.
5. **Calculate the parameter-wise absolute difference between the *best estimate SegmentLevel* and the *SegmentLevel* estimated in step 4**.

   For all subjects together:
6. **Combine the absolute difference sets for all subjects**.
   Then the distribution of this combined error may be visualised as a box plot, or its overall mean value may be estimated.

Figure 4. Algorithm used to compute the *SegmentLevel* errors for the simulated fixed-reference experiments.

The two box-plots[2] in Figure 5 illustrate the distribution of absolute difference between the *best estimate SegmentLevel* parameters, and *SegmentLevel* parameters resulting from a simulated fixed-reference experiment. In both cases, the *SegmentLevel* parameters are estimated for each subject individually, but all subjects' deviations from their best estimate are combined into a single box-plot. The distributions consist of one data point for each subject, for each sound segment, hence the relatively large range.

Figure 5 shows that when using a 1 kHz tone as fixed reference, matched with every other segment as both A/B and B/A, the resulting *SegmentLevel* parameters are typically 1.1 dB from the *best estimate* (median difference). When using a pink noise, the median absolute difference from the *best estimate SegmentLevel* is only 0.7 dB.

The notches on the box-plots show a robust estimate of the uncertainty of the medians [19]. If the notches in the boxplot do not overlap (vertically), we may conclude, at a 95% significance, that the true medians do differ[3]. Figure 5 thus indicates that using a 1 kHz tone as fixed reference leads to significantly less accurate loudness level assessments than when using pink noise.

[2] Note that the boxplots are simply used as a graphic device to illustrate the distribution of level-differences – in particular, the 'whiskers' are *not* "error bars", they merely indicate the extent of the distribution.

[3] "*A journal article that gives a good explanation of why the individual CIs are not at 95% confidence, but the "overlap test for difference of medians" is at 95% significance is Nelson, L.S., (1989), 'Evaluating Overlapping Confidence Intervals', Technometrics, 21:140-141.*" (http://www.mathworks.com/support/solutions/data/34463.shtml)

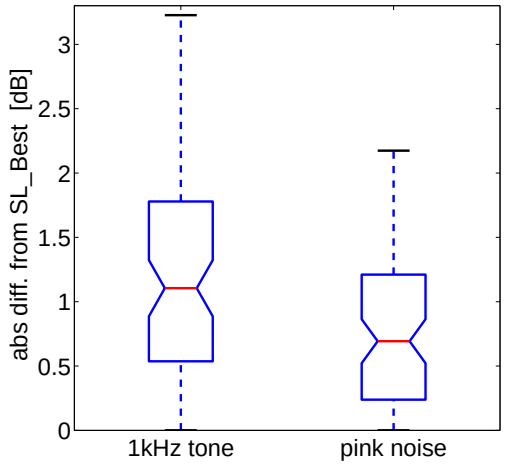

Figure 5. Distribution of the absolute difference from the *best estimate SegmentLevel* parameters, in sub-sampled experiments using either the 1 kHz pure tone or the pink noise as fixed reference.

In connection with evaluating the consequence of using test-sounds as reference, the hypothesis was posed that using pink noise or a pure tone as a fixed reference would lead to *slower* adjustments than when matching random segment pairs. Based on the experimental data, it was therefore tested whether matching music/speech against a pure tone or against pink noise was slower than matching music/speech against another music/speech segment. This hypothesis was, however, rejected ($p=0.94$): adjustment of a test-sound against music/speech was *not* significantly slower.

4.3 Speech or Music as Fixed Reference

The same procedure of sub-sampling (Figure 4) was used to investigate the consequence of using any other sound segment, from the experiment, as the fixed reference. The only difference from the test-sound investigation in the previous sub-section is that the music/speech segments were only matched against the remaining segments once (not twice).

Figure 6 shows the distributions of absolute difference from *best estimate SegmentLevel*, in sub-sampled experiments with the fixed reference segment selected from the speech, pop-rock, or classical genres. Every segment in these genres was used as fixed reference in a simulated experiment for each subject, and the results were combined according to the genre of the reference segment. Each distribution consists of one data point for each subject, for each segment in the particular genre of class.

The error distributions in Figure 6 indicate that there is no difference – on average – between using a fixed reference segment from the speech or the pop/rock genres. Choosing the reference among the classical music seems to yield slightly less accurate estimates, but this difference is barely significant.

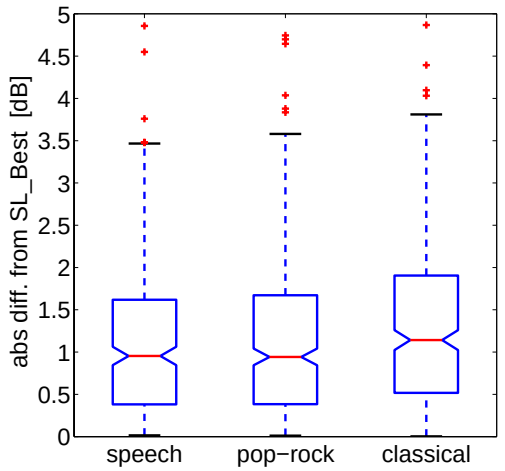

Figure 6. Distribution of the absolute difference from the best estimate SegmentLevel parameters, in sub-sampled experiments using each of the segments from the genres speech, pop-rock and classical, as fixed reference. The marks above the whiskers are *outliers*, i.e. data points beyond the maximum whisker length which is 1.5 times the interquartile range [19].

The information from Figure 5 and Figure 6 is summarized in Table 3. The median error for the pure tone is 0.42 dB higher than for the pink noise. In fact, the median error for the pop/rock segments as fixed reference, is comparable to that of the pure tone, even though the latter obtained twice as many adjustments. Recall that the speech, pop/rock, and classical entries represent the average across all segments in their respective class. The 95% percentile reveals that when using a fixed-reference experimental design, with a speech or music segment as reference, 5% of the *SegmentLevel* parameters will tend to deviate by more than 3 dB from the *best estimate*. Note that the reported differences measure how close each individual subject gets to his own *best estimate SegmentLevel*, when using a specific subset of his adjustments. If the between-subject variability had been taken into account, the deviations would have been larger.

An additional source of error is the *best estimate SegmentLevel* parameters themselves which are just – estimates. In the pilot experiment outlined in sect. 2, the standard error of the *best estimate SegmentLevel*, for the subjective loudness, were in the range 0.3 to 0.4 dB, for every individual subject.

Segment used as reference	**Matches per segment pair**	**Percentiles of absolute difference from *best estimate SegmentLevel* parameters (dB)**				
		5%	**25%**	**50%**	**75%**	**95%**
speech	1	0.06	0.38	0.95	1.62	3.00
pop-rock	1	0.06	0.38	0.94	1.67	3.05
classical	1	0.07	0.52	1.14	1.91	3.51
1 kHz tone	2	0.07	0.54	1.11	1.78	3.01
pink noise	2	0.00	0.24	0.69	1.21	1.91

Table 3. Distribution of the absolute difference from the *best estimate SegmentLevel* parameters for 5 types of (sub-sampled) experiments with a specific segment as the fixed reference to which the other segments are compared.

Table 3 seems to suggest that the most accurate results are obtained by using a pink noise segment as fixed reference. Please note, however, that the investigation only shows that the pink noise is superior to speech/music segments as fixed reference, when the noise is matched twice with every other segment whereas the speech/music reference is only matched once with the other segments. The more "fair" comparison, where the speech/music references also get two matches against every other segment was not possible to simulate because these pairs were not all part of the pilot experimental design (see section 2.1). This test could be part of a future experiment.

4.4 Balanced Pair-matching: Full vs. partial experimental designs

For a larger collection of different sound segments or stimuli, the *full* experimental design, with *nSegments*(nSegments-1)/2* segment pairs to be matched, would not be not practical. Therefore, it is relevant to investigate how the accuracy of the experiment depends on the number of pairs that are matched. Section 3 described how a balanced pair-matching experimental design may be constructed for a given number of adjustments. Spending more adjustments will undoubtedly lead to a better suppression of the experimental error in the results. But intuitively, due to the redundancy in performing on the order of $nSegments^2$ matches of *nSegments* sounds, the "last" matched pair appears to contribute less than the "first", in a *full* experiment. For a particular number of adjustments that the test subjects can perform, there is a so-called *exploration/exploitation* trade-off, as a larger number of stimuli implies a smaller number of pairs involving each stimulus, and vice versa.

This aspect of our experimental design was also investigated based on the data from the pilot experiment. The re-sampling procedure employed for this investigation is similar to the sub-sampling procedure introduced in section 4.2, except that this time numerous small balanced pair-matching experiments were constructed, which fitted inside the pilot

experiment data set. The algorithm is shown in Figure 7 – note that it is rather computationally intensive due to the high number of balanced experimental designs (step 3).

1. **Set the number of adjustments used per test subject in this set of experiments:**
   nSegments < *nSampledAdjustmentsPerSubject* < *nSegments*(nSegments-1) / 2* .

   For each test subject:
2. **Using all adjustments, compute the *best estimate* of the *SegmentLevel* parameters.**
   The *SegmentLevel* parameters are determined relative to each other; a fix-point is used to anchor them onto the phon scale.
3. **Construct a random *balanced pair-matching* experiment design**,
   with *nSampledAdjustmentsPerSubject* adjustments of distinct segment pairs.
4. **Select the subset of the subject's adjustments corresponding to the experimental design of step 3.**
   If this experimental design is not possible to simulate, with the available adjustments, repeat from step 3.
5. **Compute a new estimate of the *SegmentLevel* parameters, using only the selected adjustments.**
   Furthermore, a constant offset to all *SegmentLevel* parameters is estimated and used to align the new *SegmentLevel* parameters with the *best estimate SegmentLevel* from step 2.
6. **Calculate the parameter-wise absolute difference between the *best estimate SegmentLevel* and the *SegmentLevel* estimated in step 5**.
7. **Repeat from step 3, 500 times or more.**
   Resampling is utilized to get a good estimate of the error distribution, based on the available data. (If no 'resamples' were used at all, only the *nSampledAdjustmentsPerSubject* adjustments first selected in step 3 would influence the result.)

   For all subjects together:
8. **Combine the absolute difference sets for all subjects**.
   The distribution of this combined error may be characterized in terms of its quartiles and extrema.
9. **Repeat from step 1, with a different *nSampledAdjustmentsPerSubject* value**,
   to get an estimate of the error distribution as a function of the experiment size.

Figure 7. Algorithm used to compute the *SegmentLevel* errors for the resampling experiments.

The output of the resampling algorithm (Figure 7) was used to construct the graphs in Figure 8. The curves show how much the *SegmentLevel* estimates deviated from the *best estimate SegmentLevel* (based on the *full experiment*), by including only a smaller fraction of all the possible segment pairs. In each resample iteration of the algorithm, numerous balanced pair-matching designs, with *nSampledAdjustmentsPerSubject* segment pairs, are constructed – all fulfilling the requirements in eq. 10. (In practice, each new balanced pair-matching experimental design is constructed randomly, within its specifications, and may therefore not be unique.) Additionally, Figure 8 contains the typical absolute deviation of the 5 different genres of fixed reference, as described in section 4.3.

Figure 8 shows that, with 70 or more adjustments, the *balanced pair-matching* method will in 99% of all cases (designs) yield a better *SegmentLevel* estimate than using the *fixed-reference* method with pink noise as reference, with 40 adjustments. Note that the 70 adjustments are still considerably less than the 190 adjustments required for the *full experiment*.

It would have been fair to compare the accuracy of the *fixed-reference* results with balanced pair-matching results, based on the *same* number of adjustments. However, the performed pilot experiment did not contain the repetitions of adjustments that would have been required to evaluate the *fixed-reference* method (with speech/music references) at, for instance, *nSampledAdjustmentsPerSubject* = 40. Therefore we cannot conclude whether a speech fixed-reference design would be more or less accurate than a balanced pair-matching design, given for instance 3**nSegments* = 60 adjustments. We can conclude, however, that *without* any repetitions the fixed-reference designs are inferior to *any* balanced pair-matching design, when a certain redundancy can be afforded.

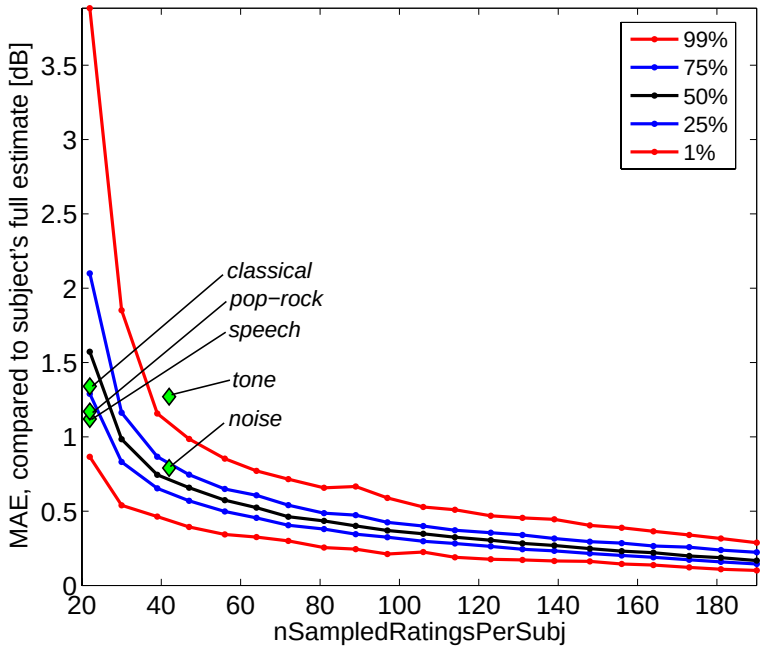

Figure 8. Mean absolute error of *SegmentLevel* parameters compared to each subject's *best estimate SegmentLevel*. In each simulated experiment, *nSampledAdjustmentsPerSubject* adjustments are used, and is resampled 500 times with different balanced pair-matching designs[4]. This procedure is repeated for each subject. The curves show the distribution of the average deviation in the simulated experiments, by its quartiles and 1% and 99% percentiles. The average deviation of the five experiments using *fixed reference* sounds are plotted at the point corresponding to the number of adjustments used in that sub-sampled experiment.

When considering the decrease of the mean deviation of the resampled pair-matching *SegmentLevel* estimates from the *best estimate SegmentLevel*, it appears that only little accuracy is gained by going beyond 80 adjustments or so (Figure 8). The *SegmentLevel* estimates do improve by obtaining adjustments of a larger fraction of the possible segment pairs, but more and more slowly. A linear regression analysis of the data indicates that the absolute deviation (*SegmentLevel* error) decreases linearly as a function of the logarithm of the number of adjustments.

The idea, that 'inside' the *full experiment* there are many smaller experiments, implies that the pilot experiment itself is a subset of some experiment with a greater redundancy – an even larger number of adjustments of the same stimuli. Hence our *best estimate SegmentLevel* parameters would themselves deviate somewhat from the *SegmentLevel* estimate of some larger experiment. Therefore the dB-values in Figure 8 are not exact, but are approximations biased by the particular *best estimate* against which the simulated experiments are judged.

4.4.1 The 50% error-reduction point

The results from the preceding sub-section show that simulated experiments with a number of loudness matches closer to the *minimum experiment* than to the *full experiment* would tend to achieve an estimate that was closer to the *best estimate* than to the estimate based on a typical *minimum experiment*. In other words, the estimates of the loudness level of the segments did improve by including matches of a larger fraction of the possible segment pairs, but they improved more and more slowly.

As a generalisation of Figure 8, the curve in Figure 9 shows the expected mean absolute deviation from a *best estimate* based on the *full experiment* design. Although we cannot predict the value of the deviation (in dB), the location of **the point of 50% error reduction** could be calculated, assuming a logarithmic relationship. This point indicates how many adjustments would be required (on average) to reduce the deviation from the *best estimate SegmentLevel* parameters, as achieved by the minimum experiment, by 50%. We shall call this number of adjustments the "50% point".

[4] The value of *nSampledAdjustmentsPerSubject* in Figure 7 starts at 22 and ends at 190, rather than starting at 20 (=*nSegments*) and continuing to the full number of adjustments performed by each subject (231). The reason for this range is that multiple random balanced pair-matching designs must be constructed at each value of *nSampledAdjustmentsPerSubject*. The *best estimate* is based on all the 231 adjustments available from each subject, that is, the 190 pairs of the *full experiment* plus repetitions of the test sounds.

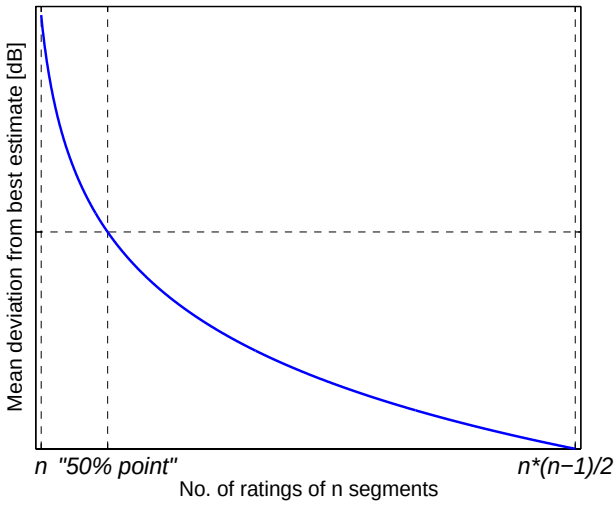

Figure 9. The graph illustrates the decreasing expected mean deviation from the *best estimate SegmentLevel*, as a function of the number of segment pairs included in a balanced pair-matching experimental design, assuming a logarithmic relationship. The "50%" point indicates the number of segment pairs required to achieve a mean deviation from the *best estimate*, that is halfway between the deviation of the *minimum experiment* and that of the *full experiment*.

Table 4 shows the consequence of using either a minimum experimental design, a balanced pair-matching experimental design at the 50% error reduction point, or a full experimental design. The number of adjustments required for the full experiment, and the time it would take to perform the experiment, is forbiddingly large when the number of sound segments greater than 100 or so. When a loudness-matching experiment is performed in psychoacoustics, to research a specific part of the perception of a certain type of sound, the number of stimuli will in some cases be fairly small, such as 20. On the other hand, when loudness-matching experiments are carried out in order to construct "subjective reference data sets", with the stimuli consisting of representative samples from multiple genres of real-world material, the number of stimuli may be in the hundreds.

Type of experimental design	*nSegments*	*nAdjustments*	Expected duration
Minimum experiment	20	20	4.7 min
Balanced pair-matching ("50%")	20	62	14 min
Full experiment	20	190	44 min
Minimum experiment	100	100	0.4 hours
Balanced pair-matching ("50%")	100	704	2.7 hours
Full experiment	100	4950	19 hours
Minimum experiment	500	500	1.9 hours
Balanced pair-matching ("50%")	500	7898	31 hours
Full experiment	500	124750	485 hours

Table 4. The number of adjustments for three different experimental designs, with 20, 100, or 500 sound segments as stimuli. The expected duration is based on the median of the effective response time, 14 s, for the subjects in the conducted loudness-matching experiments.

The minimum experiment could be, for instance, the fixed-reference method, without repetitions. In the balanced pair-matching experimental design, the number of adjustments is calculated as halfway between minimum experiment and the full experiment, on a log *nRatings* scale. In the full experiment, every segment is matched against every other segment.

4.4.2 Balanced vs. non-balanced

In an experimental design with the number of adjustments n, where $nSegments < n < nSegments*(nSegments-1)/2$, the particular n segment pairs to match could be selected as a random subset among the $nSegments*(nSegments-1)/2$ possible unique pairs, or by using a *balanced* selection. A *balanced* experimental design implies that both the absolute frequency of occurrence of each segment, and the relative frequency of occurrence of any segment compared to any other segment, are (nearly) the same for all of the segments (see section 3.2).

Thus, a minor part of the random subsets of segment pairs will by chance be balanced, and the remaining subsets will not.

Using a variant of the resampling procedure in section 4.4, the accuracy of the balanced pair-matching designs was compared to the accuracy of "non-balanced" designs with the same number of adjustments. That is, the non-balanced experimental designs would still incorporate some redundancy, but would not fulfil the requirements of eq. 10. These non-balanced designs were not completely random, however, as they –

1. consisted of distinct segment pairs, i.e. without repetitions, and
2. each (simulated) experimental design had at least one direct or indirect match between any two segments – or equivalently – the graph representing the design was *connected*. This requirement ensured that the *SegmentLevel* parameters could be uniquely estimated.

The bars in Figure 10 represent the same information as the curves in Figure 8, except that Figure 10 illustrates two different types of experimental designs. The pairs of bars in the figure show that the balanced design has a considerably lower average *SegmentLevel* error than the unbalanced designs, for relatively small numbers of adjustments. When the number of adjustments – and thus the redundancy – increases, the advantage of the balanced designs diminish. At 100 adjustments (or more), i.e., from around half the size of the full experiment, the distribution of error of the balanced and the unbalanced designs seem identical.

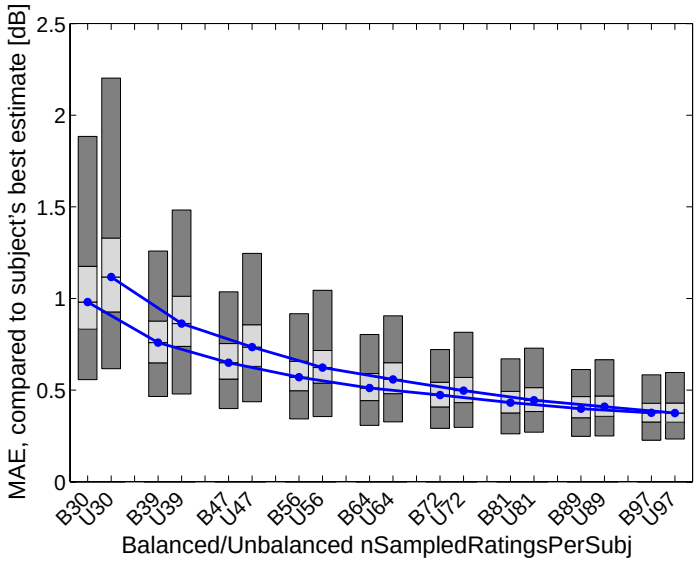

Figure 10. The distribution of mean absolute deviation of the *SegmentLevel* parameters compared to each subject's *best estimate*, using both balanced and "non-balanced" experimental designs. For example, the x-axis label "B30" means balanced pair-matching design, 30 adjustments per subject; "U" means unbalanced. The horizontal lines in each bar illustrate the error distribution's 1, 25, 50, 75, and 99% percentiles. The median deviation of the two experiment types are plotted as two curves, to illustrate their convergence.

In conclusion, the balanced pair-matching designs are most advantageous when the number of adjustments is large enough to afford some redundancy, such as the point of 50% error reduction (section 4.4.1). On the other hand, not much is gained by using a balanced experimental design, when a large redundancy can be afforded.

5. DISCUSSION AND FUTURE WORK

The presented loudness matching procedure has been designed to deliver accurate results with a limited number of matches and at the same time to enable separation of error or bias factors into separate components [1].

Now suppose that the responses of a loudness matching experiment would *not* reflect any between-listener disagreement. That is, the more adjustments or measurements each subject submitted, and the more the subject's within-listener inconsistency was evened out, the closer the results of that subject would get to the results of the other subjects. In this case, it would presumably not make much difference whether a 'minimum experiment' was conducted

with many subjects, or a 'fuller' experiment was conducted with fewer subjects. Mainly the total number of adjustments would matter. However, in [1] is was shown that, even within a homogeneous group of subjects, the between-listener disagreement was significant (albeit relatively small).

Generally, the effect of the within-listener inconsistency may be minimised through the experimental design and analysis, whereas the between-listener disagreement is a fundamental ingredient in subjective loudness assessment. This distinction between within-listener inconsistency and between-listener disagreement cannot be made, unless good loudness estimates are obtained from the individual subject – for instance by means of the balanced pair-matching experimental design presented here. The significance of subjective inconsistency and disagreement, in relation to the evaluation of *objective loudness measures*, was reviewed in [20].

As a natural continuation of the work presented here, a new loudness assessment experiment could be conducted, similar to the experiments described in this paper, but based on the experimental design with $nComparisons*nSegments^2$ adjustments, i.e., in which every segment pair is matched at least once. For *nComparisons*=1 this experiment would be roughly twice as large as the "full experiment design" described in this paper. When subsequently analysing the responses of this proposed new experiment, with the methods presented here, two advantages would be gained over the present work: 1) the results from the new experiment would provide an even better *best estimate* of the *SegmentLevel*, because it would be based on twice as many adjustments and an A/B-symmetrical experimental design, 2) the sub-sampling analyses would be able to include *any* fixed-reference design in which the reference is matched with every other segment as both A/B and B/A, and thus yield a comparison with the fixed-reference method with both 0 and 1 repetition of the adjustments.

By conducting the experiment and analysis suggested above, the assumptions presented in this paper could be tested in practice; and furthermore the results of the analysis could apply to a larger part of the experimental designs commonly used in loudness matching experiments. Nonetheless, the very same principles of resampling and sub-sampling, that have been presented here, could be applied in the new investigation.

6. CONCLUSION

The pilot experiment in a loudness assessment study was used to investigate the accuracy of certain design variations of loudness-matching experiments. Rock/pop music, classical music, speech material, and two test sounds were used as stimuli. A method based on matching all stimuli against a single *fixed reference* segment was compared to a method in which both segments of a pair were selected among all the stimuli – the *balanced pair-matching* method. In the latter method, the experimental design is constructed with redundancy, in the sense that each subject performs more than one loudness match involving each sound segment. By affording a certain redundancy and by using a balanced experimental design, the accuracy of the results for the balanced pair-matching method was improved beyond the fixed-reference method with any choice of segment as reference. The results indicate that the improved accuracy was caused partly by using the balanced pair-matching method, and partly by the limited redundancy in the design. Further experiments would be required to isolate the influence of these two factors.

The accuracy of the different experimental designs was measured as the average deviation from the *best estimate* of the loudness levels of the stimuli. A procedure of statistical resampling was developed to compute the deviation by means of simulated sub-experiments. This deviation (in dB) decreased proportionally to the logarithm of the number of adjustments included in the balanced experimental design. On the other hand, not much extra accuracy was gained by using a large redundancy, such as matching every segment with every other.

The empirical results furthermore showed that the balanced pair-matching designs significantly reduced the worst-case deviations, compared to the worst-cases of unbalanced designs. This advantage of the balanced designs was particularly pronounced when the number of adjustments was close to the number of different stimuli, i.e. with little or no redundancy, which has often been the case in previous loudness-matching experiments.

7. ACKNOWLEDGEMENTS

We would like to thank René Quesnel at the Center for Interdisciplinary Research in Music Media and Technology, McGill University, for conducting the loudness assessment experiment on which this investigation is based. Also thanks to the students in the Sound Recording program at McGill University who participated as subjects in these listening experiments.

8. REFERENCES

[1] Skovenborg, E., Quesnel, R. & Nielsen, S.H. (2004) "Loudness Assessment of Music and Speech", in Proc. of the AES 116th Convention, Berlin.
[2] Zwicker, E. & Fastl, H. (1999) "Psychacoustics: Facts and Models" (2. ed.), Springer Series in Information Sciences, 22, Berlin: Springer-Verlag.
[3] Bauer, B.B. & Torick, E.L. (1966) "Researches in Loudness Measurement", IEEE Trans.on Audio and Electroacoustics, vol.AU-14:3, pp.141-151.
[4] Jones, B.L. & Torick, E.L. (1982) "A New Loudness Indicator for Use in Broadcasting", in Preprint 1878 from the 71st AES Convention, Montreux.
[5] ITU-R (2002) "SRG-3 Status Report (2), September 2002", Document 6P/145-E,
[6] Yahoo Groups (2003) "ITU-R reflector for the SRG3 at Yahoo Groups", Internet web site: http://groups.yahoo.com/group/srg3list/.
[7] Soulodre, G.A., Lavoie, M.C. & Norcross, S.G. (2003) "The Subjective Loudness of Typical Program Material", in Proc. 115th Convention of the AES.
[8] Soulodre, G.A. (2004) "Evaluation of Objective Loudness Meters", in Proc. of the AES 116th Convention.
[9] Suokuisma, P., Zacharov, N. & Bech, S. (1998) "Multichannel level alignment, Part I: Signals and methods", in AES 105th Convention, San Francisco.
[10] Zacharov, N., Bech, S. & Suokuisma, P. (1998) "Multichannel level alignment, Part II: The influence of signals and loudspeaker placement", in AES 105th Convention, San Francisco.
[11] Aarts, R.M. (1991) "Calculation of the loudness of loudspeakers during listening tests", Journal of the Audio Engineering Society, vol.39, pp.27-38.
[12] Aarts, R.M. (1992) "A Comparison of Some Loudness Measures for Loudspeaker Listening Tests", Journal of the Audio Engineering Society, vol.40:3, pp.142-146.
[13] Poulsen, T. (2002) "Psychoacoustic Measuring Methods", Lecture note no. 3108-e, Lyngby, Denmark: Ørsted - DTU, Acoustic Technology.
[14] Griffin Technology (2003) "The PowerMate", Internet web site: http://www.griffintechnology.com/products/powermate/.
[15] Howell, D.C. (2002) "Statistical Methods for Psychology" (5. ed.), Duxbury.
[16] Trochim, W.M.K. (2004) "Research Methods Knowledge Base", Internet web site: http://www.socialresearchmethods.net/kb/.
[17] Zoubir, A.M. & Boashash, B. (1998) "The Bootstrap and its Application in Signal Processing", IEEE Signal Processing Magazine, vol.15:1, pp.56-76.
[18] Politis, D.N. (1998) "Computer Intensive Methods in Statistical Analysis", IEEE Signal Processing Magazine, vol.15:1, pp.39-55.
[19] The MathWorks (2003) "The boxplot function in MATLAB", Internet web page: http://www.mathworks.com/access/helpdesk/help/toolbox/stats/boxplot.html.
[20] Skovenborg, E. & Nielsen, S.H. (2004) "Evaluation of Different Loudness Models with Music and Speech Material", in Proc. of the AES 117th Convention, San Francisco.