Audio Engineering Society

Convention Paper

Presented at the 116th Convention
2004 May 8–11 Berlin, Germany

Loudness Assessment of Music and Speech

Esben Skovenborg[1,3], René Quesnel[2], and Søren H. Nielsen[3]

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

[2] Center for Interdisciplinary Research in Music Media and Technology,
McGill University, 555 Sherbrooke St. West, Montréal, QC, Canada H3A 1E3
rene.quesnel@mcgill.ca

[3] TC Electronic A/S, R&D Department, Sindalsvej 34, DK-8240 Risskov, Denmark
shn@tcelectronic.com

ABSTRACT

An experiment was performed to investigate the assessment of loudness of music and speech using a General Linear Model. Eight expert listeners participated in the experiment. The method of adjustment was used for loudness matching of stimuli. Both stimuli of each pair were selected from a collection of 147 homogeneous audio segments including representative samples of speech, jazz, rock/pop, and classical music, together with pink noise and a 1 kHz tone. For each segment, a reliable estimate of the loudness level was obtained from the model. Both the uncertainty and the subjectivity factors were shown to depend on the class of the stimuli. An alternative categorization based on four MPEG-7 Audio Descriptors was also used for the analysis.

1. INTRODUCTION

Loudness is a fundamental element of sound perception. Many of the factors contributing to loudness are well understood. The most prominent factor is the sound pressure level, but also the frequency content and duration of the sound influence the loudness.

The relationship between these factors and loudness have been extensively studied in classical psychoacoustics, under laboratory conditions, using relatively simple sound signals such as pulsed and continuous sine waves, and noise of various bandwidths. Such signals are typically stationary and/or synthetic, as opposed to real-world signals. Predicting the loudness of, for instance, music and speech, is not straightforward from the results of these studies.

Most everyday listening situations involve reproduced audio content which is a combination of music and speech. To complicate matters further, most of this material has been dynamically or spectrally processed, in order to fulfill aesthetical and/or technical requirements.

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* (e.g. discussed in [1]).

Objective measurement procedures of perceived loudness have been under continual development for decades. The applications for such procedures include loudness meters [2, 3] [4] [5, 6] [7, 8], and devices for loudness control [9] [10, 11] [12]. The development and evaluation of objective loudness measurement procedures are generally based on the results of loudness assessment experiments.

Measuring the perceived loudness in a listening experiment implicates numerous uncertainty factors. One challenge of designing and conducting a listening experiment is to minimize the uncertainty factors. Furthermore, the analysis of the experimental data should ideally isolate and estimate the effect of each factor. This paper describes the considerations and results of such a loudness assessment experiment.

2. EXPERIMENTAL METHOD

2.1. Choice of Method

A variety of experimental methods could potentially be used for investigating the loudness of music and speech. The method of *paired comparisons* has been used in many types of psychophysical experiments [13, 14] and consists in presenting a subject with a pair of objects or stimuli. The subject is then required to choose one of the objects based on a specified criterion. The method is easy to understand and perform, and associated statistical analyses are well-developed, e.g. [15]. An experiment on loudness assessment using this method would present the subject with pairs of sound segments and the subject's task could be to choose the loudest segment in each pair. In a variant of this method, the subject could be asked to rank the loudness of more than two stimuli presented together.

In a *scaled paired comparison* experiment, the subject must quantify the difference between two objects in a pair and express the difference on a scale of the property under investigation. This scale may be an interval scale or an ordinal scale. In such experiments the subject provides more information in each response compared to a simple (binary) selection. The statistical analysis is different from that of the simple paired comparison, [16] provides an example.

Even though many experimental studies on subjective loudness present stimuli in pairs, the method being used differs from classical paired comparisons because the property under investigation can be directly controlled by the subject. This type of control would not be possible, for example, in an experiment involving a choice between different recipes of ice-cream – it would be difficult to control the taste directly.

Loudness matching experiments [17, 18] typically use the *method of adjustment* (MOA) in which a subject is asked to adjust the loudness of a comparison stimulus using a volume or gain control until it matches a reference [3, 19, 20, 21, 22, 23]. The method of adjustment is generally reported as being intuitive and efficient, but is known to produce bias effects [24] that can however be minimized with an appropriately designed procedure. The MOA procedure also helps subjects concentrating on their task by involving them actively in the adjustment process. Lydolf [25] compared 8 different psychophysical methods in loudness experiments; he used the experiments for estimating the absolute threshold of hearing. Six of the methods were *adaptive*, meaning that the presentation level of the stimuli is automatically adjusted according to a fixed procedure. The *two-alternative forced choice* method (2AFC) has also been used in loudness matching experiments, e.g. [26], and an adaptive variant of the 2AFC in [27, 28].

The choice of a loudness matching procedure involving the method of adjustment in the experiments reported here was motivated by the following considerations:

- All subjects participating in the experiments were trained in audio engineering and thus very familiar with the task of adjusting levels using a knob.
- The method is faster than others in obtaining a quantitative rating of a given pair of stimulus, and therefore suitable for an experiment involving a relatively large number of pairs to match.
- The method is suitable for relatively long stimuli (10-15 s) because the subject may adjust the relative level and assess the loudness while listening to the stimulus.

- The responses are analyzed via a statistical model, thereby dealing with the MOA's lower accuracy of individual responses.

2.2. Setup

The experiments were conducted in the technical ear training room at McGill University. Sound stimuli were played back from a Macintosh G3 computer through a MOTU 2408 audio interface connected to a Yamaha 03D digital mixer. Two Genelec 1031A self-powered loudspeakers were placed at 30º on each side of the listener at a distance of 1.6 meters and a height of 1.2 m, as commonly found in domestic setups. Although all stimuli were monophonic, a stereo loudspeaker setup was chosen rather than a single loudspeaker since most domestic music listening is typically done on two loudspeakers.

2.3. Subjects

Eight expert listeners who had no reported hearing problems participated in the main experiment. Four of these subjects also participated in a pilot experiment. There were 6 males and 2 females, aged between 23 and 47 years old. All were enrolled in the Graduate program in Sound Recording at McGill University and were actively engaged in daily work involving critical listening, sound recording, mixing, and technical ear training.

2.4. Stimuli

Sound stimuli used for the experiment consisted of 145 monophonic segments of speech and music. The choice of monophonic stimuli was made to eliminate possible extra factors introduced by stereo signals that could have distracted the listeners from their tasks of assessing the overall loudness by directing their attention to irrelevant details of the stereo image. Such factors might affect the loudness assessment and be difficult to eliminate.

The segments were extracted from commercial music recordings, radio broadcasts, and movie soundtracks. Each segment was edited into a short excerpt of approx. 10 to 15 seconds in duration. Each segment was selected to be homogeneous with respect to its spectral content, dynamic properties, and instrumentation to facilitate the assessment of its overall loudness. The stimuli were selected from four broad but distinctive classes of sound: speech, classical music, rock/pop, and jazz. 40 segments of each of the musical categories and 25 speech segments were used. The speech segments included speech and dialog without background and speech with background music and/or environmental sounds. The musical segments included both instrumental and vocal music. Because loudness depends partly on the spectral content and dynamics of sound, the segments were selected to vary within their respective category in terms of these properties so that each class of stimuli would provide the listeners with a representative range of loudness values to assess. In addition, two test signals were included in the collection of stimuli: pink noise and a 1 kHz tone. The RMS-normalized pink noise stimulus was used to calibrate the playback at 70 dB SPL, measured with a sound level meter at the listener’s position. The use of a 1 kHz pure tone allowed us to relate the subjects’ adjustments to the phon scale. Table 1 summarizes the distribution of sound stimuli according to their class for both the pilot and the main experiments.

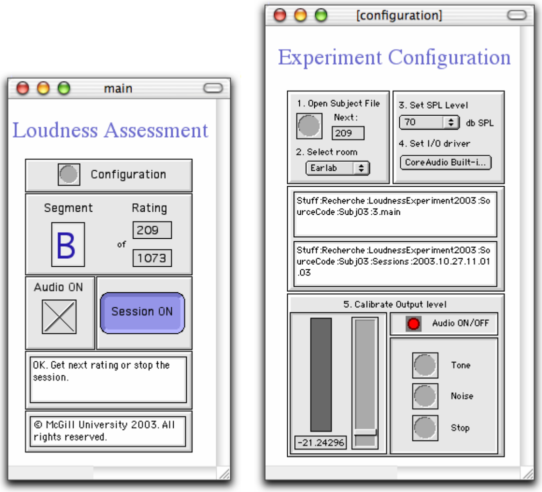
	Pilot experiment	Main experiment
pop-rock music	6	40
jazz music	0	40
classical music	6	40
speech	6	25
test sounds	2	2
Total no. of segments	20	147

Table 1. Sound segments in the pilot and main experiments

2.5. Procedure

Custom software was developed to run the experiment using the Max/MSP software tools from Cycling ’74 (see Figure 1 for screen-shots of the user interface). Subjects were presented with A/B pairs of looped audio segments. In each pair, subjects had to adjust the level of the B segment until it had the same overall loudness as the fixed-level A segment. In each pair, both segments were chosen from the collection of 147 stimuli and could consist of any combination of sound classes (e.g., jazz-jazz, speech-pop, classical-rock, etc.).

Subjects had an initial training session during which they became familiar with the procedure and software. At the beginning of each subsequent listening session, they had a short “warm-up” period during which they could rate a few practice pairs. There were no pre-determined number of ratings to do during any given listening session but the maximum duration of a session was set to 1 hour to avoid listening fatigue. Subjects could pause the experiment at any time, mute the sound, and rest if they wished to. The software allowed a session to be interrupted and resumed later with the next stimuli pair to be rated. Detection of user errors was built into the software to prevent the same stimuli pair to be matched twice and to avoid skipping of a match. For each adjustment, the software recorded the subjects' level setting (in dB), their response time in milliseconds, and the number of A/B comparisons.



Rating 209 of 1073

OK. Get next rating or stop the session.

© McGill University 2003. All rights reserved.

1. Open Subject File Next: 209

2. Select room Earlab

3. Set SPL Level 70 db SPL

4. Set I/O driver CoreAudio Built-i...

Stuff:Recherche:LoudnessExperiment2003:SourceCode:Subj03:3.main

Stuff:Recherche:LoudnessExperiment2003:SourceCode:Subj03:Sessions:2003.10.27.11.01.03

5. Calibrate Output level -21.24296

Figure 1. Screenshot of the Max program that was developed.

The typical sequence of events during a listening session was straightforward (Figure 2). At the beginning of each session, the subject's data file and playlist were loaded and playback level was calibrated to 70 dB SPL. Then the subject requested each sound pair to rate. The subject could compare at will between the A (reference) and B stimulus, adjust the level of B until it was judged that both had the same overall loudness. When an adjustment for a given pair was registered, the sound was turned off and the next stimuli pair was subsequently loaded at the subject's request or the subject could choose to terminate the session.

An external volume knob was used to make loudness adjustments (Figure 3). The choice of a real rotary knob to adjust the level of the stimuli offered significant advantages over the use of the mouse and an onscreen graphic control such as a slider or a dial. The knob had no endpoints and did not provide tactile nor visual feedback to subjects about its quantization level and its neutral position. Thus, for each adjustment, subjects were not biased by the knob's previous position and could base their judgment solely on what they heard. It had a natural feel to the subjects and proved more intuitive to use than clicking and dragging a mouse. Its resolution was set to 0.25 dB with a range of ±24 dB.

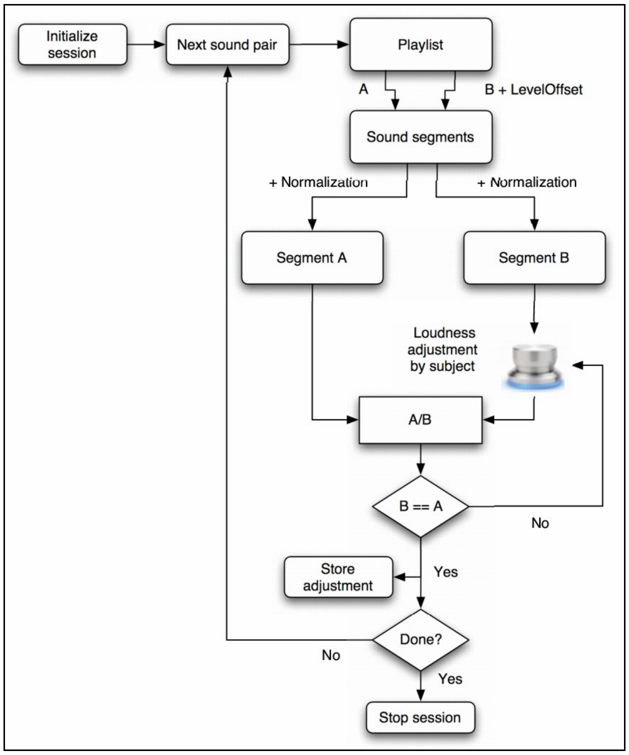


Figure 2. The experiment procedure.

Figure 3. The "Power Mate" knob used by the subjects to control the relative level.

2.6. Level normalization and spread

A two-step procedure was applied to control the spread in level of the sound segments. In step 1, each segment was first filtered to produce a B-weighting curve, which approximates the frequency sensitivity of the human ear at medium SPL levels [29]. The B-weighting was preferred to the more common A- or C-weightings in order to correspond to the 70 dB SPL presentation level used in the experiments reported in the present paper. The level of the segment was then normalized by calculating the gain required to make its overall RMS level equal to that of the pink noise segment used for SPL calibration. This B-weighted RMS normalization step produced a rough loudness equalization that aligned all segments in the same range as the calibration sound. Figure 4 shows the distribution of normalization gain applied to the segment collection (Table 1), and is not atypical for a collection composed of segments from various sources.

In step 2, a random level offset, drawn from a uniform distribution of random numbers in the interval [–6..6] dB, was added to the B segment of each A/B segment pair. The purpose of this offset was twofold. First, it increased the spread of levels, thus increasing the average level adjustment required by the subject. This insured that the adjustments performed by subjects would be larger than the loudness JND and larger than the level quantization step used in the experiment. It also minimized cases for which no level adjustment would be required from the subject.

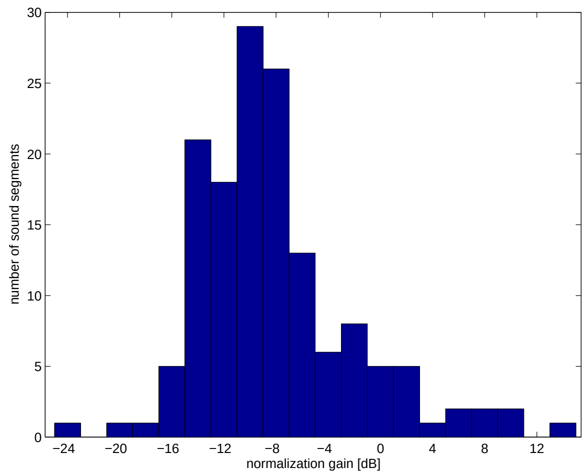


Figure 4. Normalization gain applied to the 147 sound segments used as stimuli.

Second, the random level offset was also used to scramble the segment order implied by the normalization in the previous step. The normalization was *not* a perfect loudness equalizer, therefore some segments were louder, after the level normalization, and others were softer. Without a level offset, the loudest segment, for instance, would have always required a negative level adjustment whenever it was presented as the B segment in a pair. This could have caused a bias in the adjustment of the individual segments. By applying the random level offset, the loudest segment (after normalization) was no longer always the loudest segment in each pair.

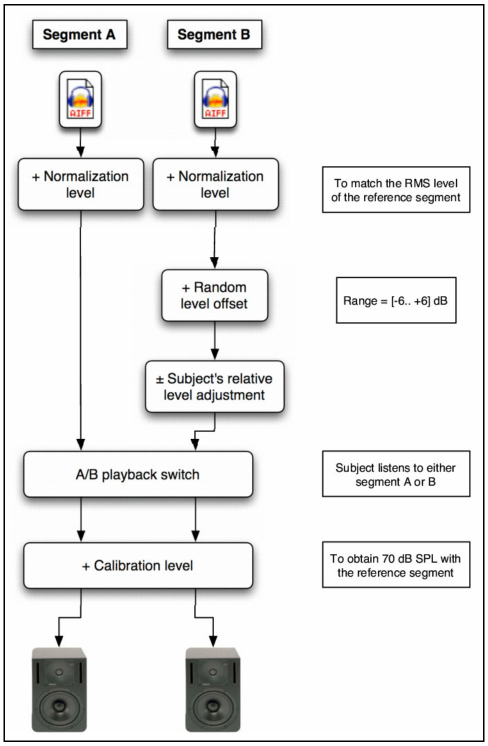

Segment A
Segment B
AIFF
AIFF
+ Normalization level
+ Normalization level
To match the RMS level of the reference segment
+ Random level offset
Range = [-6.. +6] dB
± Subject's relative level adjustment
A/B playback switch
Subject listens to either segment A or B
+ Calibration level
To obtain 70 dB SPL with the reference segment

Figure 5. The signal path with the different controls of the level.

As illustrated in Figure 5, the playback level for each stimulus depended on –

1. the normalization level specific to the segment, calculated from the B-weighted RMS level of the sound file,
2. the random level offset, specific to the B segment of each pair to be matched,
3. the subject's adjustment of relative level of the B segment, for each pair to be matched, and
4. the calibration level of the Max software (used to match the digital level to the desired SPL).

2.7. The balanced pair-matching method

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, subjects match all stimuli against a single sound segment selected in advance. When using the *balanced pair-matching* method, both segments in a pair are drawn from the same collection. The composition of the set of pairs to be matched is said to be *balanced* because the frequency of occurrence of the different segments is the same.

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,
2. the bias due to the subjective impression of the particular fixed reference segment is avoided (or at least spread out over all segments),
3. all the obtained level adjustments are used in the estimate of the loudness of every segment (via the model described in section 4), as opposed to using only the fraction *1/nSegments* of the adjustments for estimating the loudness of each segment, when using a fixed-reference scheme.

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 experiment design* requires only *nSegments* adjustments (e.g., each segment *i* could be matched with segment *i+1*). Thus the *redundancy* in the experiment 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}*/2* different segment pairs.

Prior to the main experiment, a pilot experiment was carried out using half the number of subjects and a scaled-down collection of sound segments. In the pilot experiment, the effect of using the fixed-reference method, compared to using the balanced pair-matching method with various degrees of redundancy, was investigated with a technique inspired by resampling statistics. Suppose that the *best estimate* of the loudness level of every segment is the estimate based on the responses from the full experiment, i.e. all adjustments made by a given subject. Briefly, the resampling technique consists of computing the deviation from the *best estimate*, given a subset of the available responses that corresponds to some smaller experiment design. This procedure is repeated many times, with different experiment designs of various sizes. Note that the deviation from the *best estimate* is calculated based on each individual subject's adjustments only; hence, the between-subject variability is not taken into account. The details of this investigation is beyond the scope of this paper and will be the topic of a subsequent paper.

In the pilot, every subject completed the *full experiment design*, i.e., each subject adjusted the relative loudness of every segment against every other segment. The matching sequences of the pairs were randomized to suppress the effect of any factors related to the timeline of the experiment. Because the number of different segment pairs for a full experiment is nearly *nSegments*^{2}*/2*, only a limited number of segments could be included in the pilot. In addition to the two test-sound segments, 6 segments were selected in each of the classes, pop/rock, classical music, and speech, to constitute a sub-sample of representative segments from each class.

Because every segment was matched with every other segment in the pilot, the *fixed-reference* type of experiments could be simulated by selecting the subset of the responses in which a chosen reference sound segment is compared to every other segment. Based on this kind of subsampling of the pilot experiment responses, the consequence of using different sound segments as the fixed reference was examined. The distributions of absolute difference from the *best estimate* indicate that there is little or no difference – on average – between using a fixed reference selected from the speech, pop-rock, or classical music classes. No single best fixed-reference segment could be identified. Furthermore, a balanced pair-matching design with even a small redundancy leads to a better estimate than using the best of the fixed-reference sub-experiments.

For a larger collection of segments or stimuli, the full experiment design with *nSegments*(nSegments-1)/2* matches is not practical. Therefore it is relevant to investigate how the quality of the experiment depends on the number of pairs that are matched. Using more adjustments will undoubtedly lead to a model with better suppression of the experimental error. 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". For a given number of matches that the test subjects can perform, there is a so-called exploration/exploitation trade-off, as a larger number of sound stimuli implies a smaller number of pairs involving each stimulus, and vice versa.

This aspect of our experimental design was also investigated using the pilot data and the resampling procedure. Numerous small balanced pair-matching experiments, which fitted inside the pilot experiment data set were constructed. The results indicated that (sub-)experiments with a number of loudness matches much closer to the *minimum* than to the *full experiment* tended 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 of the segments did improve by making use of matches of a larger fraction of the possible segment pairs, but they improved more and more slowly. The data suggest that the deviation from the *best estimate* decreases linearly as a function of log. number of matched pairs.

In an experiment design with the number of matches *n*, where *nSegments < n < nSegments*(nSegments-1)/2*, the particular *n* segment pairs can be selected as a random subset among the *nSegments*(nSegments-1)/2* possible pairs, or by using a *balanced* selection. A *balanced* experiment 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. Additionally, the A/B order and the direction of the adjustment level need to be balanced to best counteract bias phenomena. Thus, some of the random subsets of segment pairs will be balanced, and some subsets will not. Results from the pilot experiment showed that the balanced pair-matching designs significantly reduced the worst-case estimate compared to the worst-case unbalanced (random-subset) designs. However, this advantage of the balanced designs decreased as they approached the size of the full experiment.

In conclusion, the balanced pair-matching experimental designs are most advantageous when the number of matched pairs is large enough to afford some degree of redundancy. On the other hand, not much extra accuracy is gained by using a large redundancy. The fixed-reference experiment designs (using *nSegments* matches) is inferior to *any* balanced pair-matching design, when a certain redundancy can be afforded. For the pilot, using 2-3 times *nSegments* as the number of matches per subject was sufficient to obtain the improved accuracy.

3. RESULTS

The pilot experiment was employed partly to verify the functionality of overall procedure, the setup and software, and partly to test certain hypotheses regarding the experimental method (see section 2.7). The following sections are all be based on the results from the main experiment. Section 4 will present the experiment results in the context of a statistical model.

3.1. Number of Ratings, Range of Levels, and Performance of Test Subjects

The details of the pilot and main experiments are listed in Table 2. The *response time* is the time during which the subject is listening to and adjusting the relative level of the segments, so the time spent calibrating the playback level, subject's warming up, starting the session etc. is not included. In the main experiment, the median response time of the subjects is 14 seconds, which seems efficient considering that the combined duration of the A and B segments of each pair was 20 to 30 seconds. Presumably the subjects could adjust the relative level faster than the combined duration of the segments because they would trust the homogeneity of the segment. On average, subjects switched between the A and B segments 6 times for each loudness match.

The *AdjustmentLevel* variable contains the actual adjustments of relative level, and the *DifferenceLevel* variable is the *AdjustmentLevel* with the random level offset removed. In the main experiment, the *AdjustmentLevel* had a standard deviation of 4.2 dB, whereas the standard deviation of the *DifferenceLevel* was 2.9 dB; the difference is caused by the extra variance added via the random level offset. The total range of the *DifferenceLevel* is 27 dB (Figure 6). These statistics are all measures of the controlled *dynamic range* of the experiment (cf. section 2.6).

Note that *if* the level normalization had employed a (hypothetical) ideal loudness function, then the optimal adjustment for every pair of segments would have always been *DifferenceLevel* = 0 dB. In this case, the variability of the *DifferenceLevel* would have comprised only experimental error, such as adjustment bias and within-listener inconsistency, and between-subject disagreement (what could be called "the subjective factor of loudness"). However, only a primitive level normalization was applied. Hence the variability of the *DifferenceLevel* additionally comprises the part of the relative level which the subjects agree on, i.e. the "common loudness".

	Pilot experiment	Main experiment
Total number of sound segments (*nSegments*)	20	147
Number of test subjects	4	8
Total effective response-time, all subjects	3.8 hours	37.8 hours
Median response-time	12.3 seconds	14.2 seconds
Median number of A/B-comparisons	5	6
Number of all possible segment-pairs, with A ≠ B, *nSegments* (nSegments-1)/2*	190 pairs	10731 pairs
Percentage of all possible segment-pairs that are rated by each subject	100 %	10 %
Number of ratings or adjustments submitted per subject[1]	231	1073
Average total number of ratings or adjustments per subject of each segment as A or B	23.1	14.6
Average total response-time, per subject, per segment	2.85 minutes	1.93 minutes

Table 2. Details of the pilot and main experiments

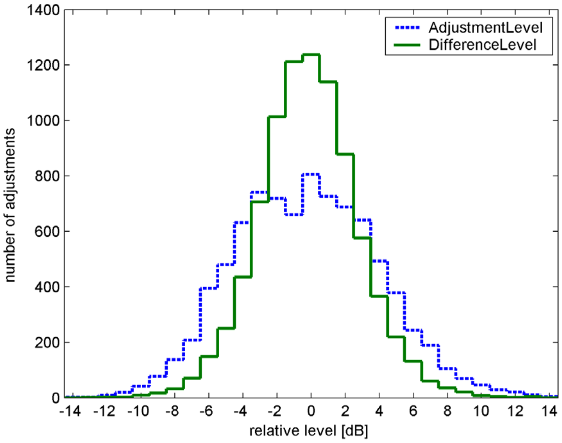


Figure 6. Histogram of the *AdjustmentLevel* and *DifferenceLevel* variables (main experiment).

3.2. Level Differences within Same-pair Adjustments

Each segment pair included in the main experiment was rated once by four different subjects – twice in each A/B order of the segment. Calculating the pairwise differences between the multiple adjustments of the same segment pair provides a simple way of assessing the between-subject agreement. This measure is also known as *Gini's mean difference* or *GMD*. The *GMD* is defined as a measure of the pairwise mean absolute difference between adjustments of the same segment pair, in either A/B order:

$$GMD(\{i,j\}) = \frac{1}{nRatingsPerPair \cdot (nRatingsPerPair - 1)} \cdot \sum_{SubjX \neq SubjY} \left| DL_{SubjX}(\{i,j\}) - DL_{SubjY}(\{i,j\}) \right| \quad (1)$$

In eq. 1 the variable DL_{SubjX} denotes the *DifferenceLevel* corresponding to an adjustment made by subject *X*. The value of the term $\{i,j\}$ is assumed to be the same for both (i,j) and (j,i), hence disregarding the A/B order of the segments in a pair. A subscript may be used to indicate the number of pairs averaged over, as in GMD_4, for the case where the *GMD* is calculated for the same-pair ratings from 4 different subjects.

3.2.1. Typical same-pair difference

The *typical* level difference in same-pair adjustments can be determined by considering the quartile statistics of the distribution of the measure of between-subject disagreement, the GMD_4. In the main experiment, when two randomly selected subjects adjusted the same randomly selected segment pair (disregarding the A/B-order), their adjustments differed by less than 1.46 dB in 25% of the cases, less than 2.08 dB in 50% of the cases (the median disagreement), by less than 2.78 dB in 75% of the cases, and by less than 4.20 dB in 95% of the cases. Please note that the *GMD* measure contains both subject-specific bias, the experimental error, and also loudness perception differences; all three factors will contribute to the measured disagreement. By using a statistical model of the data, these factors can be separated (presented in section 4).

The *GMD* was chosen as a measure of disagreement within same-pair ratings, although the *standard deviation* of the *DifferenceLevel* within each pair could alternatively have been used. However, the *GMD* was judged to be a more natural measure because it directly

[1] In the pilot, the number of adjustments made by each subject is larger than the total number of segment pairs because certain pairs were matched both as (i,j) and (j,i).

calculates the difference between the *DifferenceLevel* values from the subjects, whereas the standard deviation is inevitably based on the assumption that the different subjects' *DifferenceLevel* values are (symmetrically) spread around their mean value. Yet, the GMD_4 and the standard deviation were, for the main-experiment data, highly correlated ($r = 0.994$), so any conclusions for the *GMD* variable are likely to hold for the disagreement measure based on standard deviation as well. The median of the standard deviation of all groups of same-pair adjustments was 1.71 dB.

3.2.2. Worst case same-pair difference

What material did the subjects disagree on the most? This question can be answered by studying which segments occupy the upper tail of the *GMD* distribution – that is, a qualitative investigation of where the pairwise mean difference in *DifferenceLevel* is highest.

Roughly, the 50 segment-pairs with the highest GMD_4 seemed to account for the largest values, so the segments in these pairs should point to some sources of between-subject disagreement. The most frequently occurring individual segments in this *GMD* top-50, as either segment A or B, are listed below:

1. The 1 kHz pure tone (*class test-sound, ID = 146*) in 24% of *GMD* top-50 pairs.
   The pure tone's (infinitely) narrow bandwidth and lack of dynamics might made it very difficult to compare with music and speech segments. Another possible explanation is that loudness of the tone was particularly sensitive to the comb filtering that may occur when using a stereo speaker setup with a mono sound.
2. A very angry woman shouting "shut up..." (*class speech, ID = 8*) in 12% of *GMD* top-50 pairs.
   This segment was found "distracting" by several test subjects, hence the annoyance factor might have affected the loudness judgment. The bandwidth of the segment's spectrum was also narrow.
3. A capella female choir (*class classical, ID = 78*) in 8% of *GMD* top-50 pairs.
   This piece was spectrally atypical, with a high and narrow spectral bandwidth (again).

The remaining segments in the top-50 each occurred in less than 6% of the Top-50 pairs.

One might consider whether the *minimum GMD* could similarly point to segments leading to a *low* between-subject disagreement. However, the lowest part of the *GMD* distribution was occupied by many different segment pairs – no small number of particular segments could be identified as was the case for the maximum *GMD* corresponding to the "heavy upper tail" in the *GMD* distribution.

3.2.3. GMD in terms of Stimulus Class

The previous section examined the *GMD* in terms of specific segment pairs. The relation of the *GMD* and the different classes of the segment pairs was also investigated. Associated with each value of the *GMD* (eq. 1) is a pair of classes to which the segments A and B belong.

In Figure 7 the GMD_4 distribution is displayed for each possible pair of segment classes (the pair test-sound~test-sound was not included in the main experiment). Overall, the pairs of classes appeared to belong to two groups, in terms of mean disagreement: 1) music/speech ~ music/speech, and 2) test-sound ~ music/speech. The GMD_4 for the latter group was considerably higher, with the Q3 quartile being 2-4 dB higher than the other groups. The previous section pointed towards the 1 kHz pure tone as the source for the high disagreement associated with the test-sound class. But apparently, it was somewhat less difficult to compare the tone to speech segments than to compare it to the music segments.

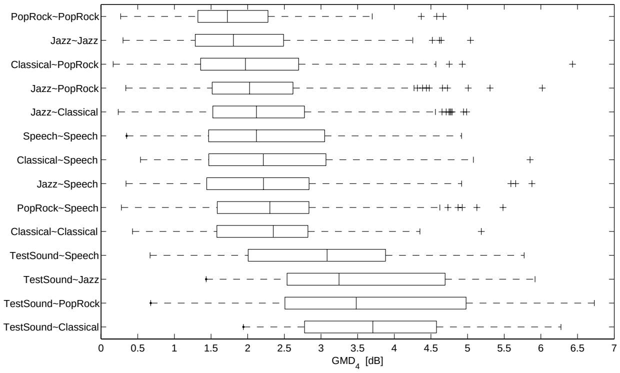


Figure 7. Each boxplot illustrates the distribution of GMD_4 for a distinct pair of stimulus classes: The 'box' indicates the quartiles and the 'whiskers' indicate the range of the distribution; each '+' is an outlier. The pairs of classes are ordered according to their median GMD_4.

The mean disagreement was lowest, typically with a GMD_4 of around 1.5-3 dB, for pairs of classes consisting of *identical* music/speech classes, with pop-rock ~ pop-rock pairs meeting with least disagreement. In general, the between-listener disagreement was higher for pairs including a speech segment, than for pairs consisting of segments from two music genres. In particular, pairs consisting of two speech segments had a higher upper quartile of disagreement than any of the music-only pairs.

An ANOVA was performed to test whether the observed difference in *GMD* among the classes was significant. The variable *ClassPairIdx* has a unique value for each different pair of segment classes, disregarding the A/B-order. The ANOVA for GMD_4 as dependent variable and the *ClassPairIdx* as main factor produced: F(13,2131) = 10.32, p<0.0001. Therefore, the observed difference was indeed significant.

4. ANALYSIS AND MODELING

Section 3 presented various analyses based directly on the responses from the main experiment. The following sections present the results from the experiment in the context of a statistical model.

4.1. A Linear Model

Each *DifferenceLevel* (*DL*) value, calculated from a response obtained in the listening experiment, is associated with the pair of audio segments that were used as stimuli. *DL* is in dB, and is calculated by subtracting the random level offset from the relative level adjusted by the subject (see section 2.6). Suppose that, underlying the set of *DL* values, there is a loudness level for each single segment, and call it the *SegmentLevel* (*SL*). The variable *DL*(*i,j*) denotes the *DifferenceLevel* for the 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, *DL*(3,5) = 2.5 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 +2.5 dB.

Let *SL(i)* denote the *SegmentLevel* value of the segment with index *i*. The relationship between *DL* and *SL* could then be expressed, as follows:

$$DL(i,j) = SL(i) - SL(j) \tag{2}$$

Eq. 2 implies that two properties should hold for the *DifferenceLevel* values obtained in the experiment. These properties are expressed in eq. 3 and eq. 4.

$$DL(i,j) = -DL(j,i) \tag{3}$$

The symmetry property above seems reasonable, considering that the stimuli A and B were drawn randomly from the same collection, and treated identically except that the subject could adjust the level of B and not of A. The symmetry also seems intuitively sound – if segment *i* is X dB louder than segment *j*, then segment *j* must be X dB softer than segment *i*. Due to bias phenomena that are investigated and quantified in section 4.3, the symmetry property in eq. 3 is sometimes not fulfilled exactly, although the discrepancy turns out to be relatively small.

$$DL(i,j) = DL(i,k) + DL(k,j) \tag{4}$$

The implication of the linearity reflected in eq. 4 is that loudness is a linear function of the SPL. As the sensitivity to different frequencies – and hence the subjective loudness – depends on the absolute SPL of the stimulus, this property is known not to be accurate. Nevertheless, within a relatively narrow SPL range, as in these experiments, such deviations from linearity are small. The transitivity of *DL* was tested positive, based on the responses from the pilot experiment (see Appendix).

All the *DifferenceLevel* values resulting from the main experiment can be considered as a set of equations of the form expressed by eq. 2. Specifically, we get *nTotalAdjustments* equations with *nSegments* unknowns (eq. 5).

$$\begin{bmatrix} DL(1,2) = 1\cdot SL(1) - 1\cdot SL(2) + 0\cdot SL(3) + \ldots + 0\cdot SL(nSegm) \\ DL(1,3) = 1\cdot SL(1) + 0\cdot SL(2) - 1\cdot SL(3) + \ldots + 0\cdot SL(nSegm) \\ DL(2,3) = 0\cdot SL(1) + 1\cdot SL(2) - 1\cdot SL(3) + \ldots + 0\cdot SL(nSegm) \\ \ldots . \end{bmatrix} \tag{5}$$

Corresponding to a *multiple linear regression* problem, an estimate of each *SL* value can be computed by established numerical methods yielding a *least-squares error* solution. Figure 8 illustrates the process of estimating the *SL* values based on the responses from the listening experiment.

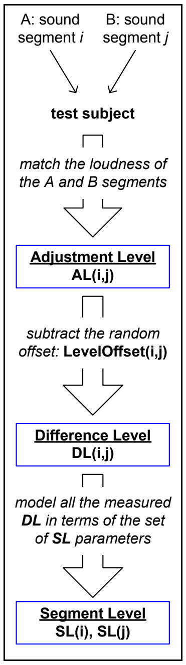

A: sound segment *i*
B: sound segment *j*
test subject
match the loudness of the A and B segments
Adjustment Level
AL(i,j)
subtract the random offset: LevelOffset(i,j)
Difference Level
DL(i,j)
model all the measured DL in terms of the set of SL parameters
Segment Level
SL(i), SL(j)

Figure 8. From a loudness match and *AdjustmentLevel*, via a model of *DifferenceLevel*, to the *SegmentLevel* estimate.

4.2. The GLM: Regression and ANCOVA

The data from the main experiment was modeled 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) ([30] [31]). The *regression* refers to the process of estimating the optimal *SegmentLevel* (*SL*) parameters, given the *DifferenceLevel* (*DL*) adjustments from the listening experiment. The *ANCOVA* is a type of analysis of variance which is characterized by containing both categorical and continuous predictor variables (as opposed to classical ANOVA which contains only categorical predictors); the continuous predictors are the so-called *covariates* of the ANCOVA. Certain statistical hypotheses regarding the significance of the different models and their parameters or factors can be tested by means of the ANCOVA.

In the following sections six different models of increasing complexity and explanatory power are introduced and evaluated in the framework of the GLM. Eq. 6 specifies the simplest of these models: each *observation* (*DL*) is explained (only) by the difference between two continuous predictor variables (*SL*), corresponding to the two segments in the matched segment pair. The residual error (ε) is a random variable which is normally distributed with a mean value of 0.

$$DL(i,j) = SL(i) - SL(j) + \varepsilon \tag{6}$$

The ANCOVA of the model, displayed in Table 3, shows that *SL* is a highly significant set of factors in predicting the observed *DL*. That is, the loudness adjustments performed by the subjects *did* depend on the pair of segments that were matched. The coefficient of determination ($R^2 = 0.52$) measures the amount of the total variance of the response variable that is "explained" by the predictor variables.

Source	*df*	*S.S.*	*M.S.*	*F*	*Pr > F*
SL	146	38002	260.3	66.26	<.0001
Residual	8438	34717	4.11		
Total	8584	72719			

Table 3. Analysis of covariance for the model of *DL* in terms of *SL*.

In the ANCOVA table, the 146 degrees of freedom (*df*) for the model correspond to 147–1 *df*, corresponding to the 147 different segments in the experiment minus one *df* as the segments are only specified relative to each other (i.e., in pairs). The total 8584 *df* is the total number of adjustments performed by all subjects, excluding a few obvious outliers.

4.3. Bias Phenomena

Some of the variance in the *DL* variable that is not explained by the *SL* factors, in the previous section, is due to a random "measurement error" which is inevitably a part of any experiment. However, the level adjustments by a subject might have been affected by some (non-random) factor *other* than the specific pair of segments to be matched. Such factors would have led to a systematic error in the subject's response – i.e. the *DL* would have been biased. It is fundamental to experimental setup and design to either eliminate or – if elimination is not possible – to randomize such factors. A factor is randomized in the hope that its effect on the response will be a random error rather than leading to a systematic error or bias [32].

When the response of an experiment is biased it may be possible to include a bias factor in the model of the experimental data. Thus the magnitude of the bias can be estimated, and its influence on the other model parameters can be eliminated or reduced. In the following sections the basic model of eq. 6 is extended with two types of bias: the A/B-order bias and the adjustment-bias.

4.3.1. A/B-order bias

One kind of bias is related to the A/B-order of the segments in a segment pair to be matched in loudness. This bias phenomenon might be caused by the fact that the A stimulus was always presented before the subject could switch to the B stimulus. This cause, however, seems unlikely to have an effect on an adjustment submitted on average 14 seconds later. A more likely cause could be related to the adjustment strategy of the subject. For example, suppose the subject would always make the last fine-tuning adjustment as an *upward* turn of the knob. This 'habit' would be a likely cause of an A/B-order bias.

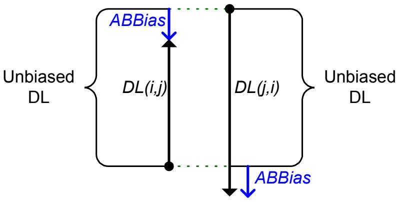


Figure 9. Illustration of A/B-order bias. The figure shows the biased adjustment of a segment pair, matched both as A/B and B/A. Left side, the biased *DifferenceLevel* (*DL*) is too small, due to a negative bias; right side, *DL* is too large due to the same bias.

The A/B-order bias corresponds to a constant term or a 'linear intercept' in the GLM model. Subjects that have a non-zero A/B-order bias will make a *systematic error* in submitting responses that are all – on average – either too high or too low (eq. 7).

$$DL(i,j) = SL(i) - SL(j) + ABBias \qquad (7)$$

Note that eq. 7 implies that for two *DL* adjustments of the same pair, done as A/B and B/A, the expected difference would be two times the estimated bias value:

$$DL(i,j) = -DL(j,i) + 2 \cdot ABBias \qquad (8)$$

If the design of an experiment was perfectly balanced, in the sense that every stimulus-pair (*i,j*) was also presented to the same subject as (*j,i*), then any A/B-order bias would be canceled out and would have no effect on the *SL*-estimates. However, the variance or uncertainty estimates of the *SL* variables would still be higher than without the bias, unless the A/B-order bias is explicitly included in the model of the data.

4.3.2. Over-adjustment bias

Another type of bias is called *adjustment bias* (*AdjBias*). The adjustment bias is caused by a systematic error related to the overall *direction* of the adjustment made by the subject – i.e. whether the submitted *AdjustmentLevel* was positive or negative. During the typical 14 seconds spent by the subject on an adjustment, it can be hypothesized that the final adjustment was made by going a little above the loudness matching level, and then a little below, and so on until the subject was satisfied and submitted the adjustment. The adjustment bias could have thus been caused either by an over-adjustment or by an under-adjustment. In the case of an over-adjustment, the subject would have adjusted the level 'too far'.

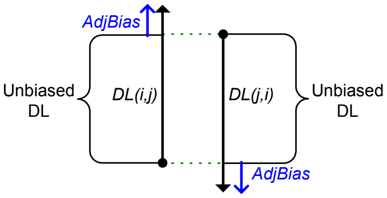


Figure 10. Illustration of adjustment bias. The figure shows the biased adjustment of a segment pair, matched both as A/B and B/A. Left side, the biased *DifferenceLevel* (*DL*) is too high, due to the over-adjustment bias; right side, *DL* is too low due to the same bias.

Eq. 9 shows how the adjustment bias is included in the model. The term *sign(AdjLevel)* has a value of 1 when the *AdjustmentLevel* is positive, a value of -1 when the *AdjustmentLevel* is negative, and 0 when the *AdjustmentLevel* (rarely) is 0 dB.

$$DL(i,j) = SL(i) - SL(j) + sign(AdjLevel) \cdot AdjBias \qquad (9)$$

4.3.3. Results of model with biases

The GLM including both types of bias, with the bias terms shared by all subjects, is given in eq. 10.

$$\begin{aligned} DL(i,j) = SL(i) - SL(j) + \\ sign(AdjLevel) \cdot AdjBias + ABbias + \varepsilon \end{aligned} \qquad (10)$$

When specifying the bias factors in the model, their values are estimated together with the *SL*-parameters.

Thus the resulting *SL*-parameter estimates will be corrected for the bias effects, unlike the estimates of the simpler model. The ANCOVA for the model with bias terms is shown below.

Source	*DF*	*S.S.*	*M.S.*	*F*	*Pr > F*
SL	146	38002	260.3	66.99	<.0001
ABBias	1	183.5	183.5	47.23	<.0001
AdjBias	1	1755	1755	451.7	<.0001
Residual	8436	32778	3.88		
Total	8584	72719			

Table 4. ANCOVA for the model of *SL* and two bias terms.

The *F*-test in Table 4 shows that both types of bias are significant; that is, the biases contribute to modeling or 'predicting' the responses from the main experiment.

Parameter	*Estimate*	*Std.Err.*	*t*	*Pr > |t|*
ABBias	-0.13	0.021	-5.96	<.0001
AdjBias	0.50	0.024	21.25	<.0001

Table 5. Estimates of the bias terms.

The estimated values for the two bias terms are shown in Table 5. Also shown is the *t*-test which confirms that they are both significantly different from a value of 0. The standard error can be regarded as a measure of the uncertainty with which the bias values are estimated, i.e. the accuracy is on the order of 0.02 dB. The largest bias term is the adjustment bias of 0.5 dB. The value is positive which means that the subjects were over-adjusting.

In the present experiments, the loudness levels corresponding to each stimulus, the *SL*-parameters, are obtained via a statistical model of the ratings. Certain other experiment types can obtain unbiased responses directly. These procedures are, however, typically less efficient than the method of adjustment employed here (see section 2.1).

4.3.4. Individual Bias

As both the A/B-order bias and the adjustment bias may well be related to the adjustment strategies of the individual subjects, the next extension of the GLM will include *individual* bias terms. This model is shown in eq. 11, where the subscript *S* implies that each subject gets his or her own two bias parameters. As each response (*DifferenceLevel*) belongs to a specific subject, only the bias terms of that particular subject will be modeling that response. However, as the adjustments of all subjects are combined, the bias terms of all subjects are estimated together, along with the common *SL*-parameters.

$$\begin{aligned} DL(i,j) &= SL(i) - SL(j) + \\ &\quad sign(AdjLevel) \cdot AdjBias_S + ABbias_S + \varepsilon \end{aligned} \tag{11}$$

The GLM includes, for each subject, a *t*-test of whether the individual bias terms are non-zero with statistical significance. The results show that -

- the A/B-order bias is significant for 6 out of 8 subjects, with values in the range of [–0.50, 0.23] dB,
- the adjustment bias is significant for 6 subjects out of 8, with values in the range of [0.29, 1.32] dB.

It is interesting to note that both types of bias are significant for 75% of the subjects, whereas the bias estimates for the remaining 25% of the subjects cannot be considered significantly as non-zero. The 75% subjects with significant bias are *not* the same for the two bias types. These findings support the assumption that these types of error are caused by *individual* habits in performing the level adjustments. The magnitude of the individual adjustment bias is in the interval 0.29 to 1.32 dB. This implies that all of the subjects with the systematic error were over-adjusting (none were under-adjusting).

It is not surprising that there is a significant individual A/B-order bias because the 'shared' bias term in section 4.3.3 was also significant. Yet it is interesting to note that the range of the bias estimates, –0.5 to 0.2 dB, indicates that some subjects (2) had a positive bias value while others (4) had a negative value. It is possible that this disparity was caused by the listeners using different listening or adjustment strategies when making their level adjustments.

In conclusion, 75% of the test subjects were making a systematic error depending on the A/B-order of the segments, and 75% were over-adjusting. The magnitude of this error depended on both the specific test subject and on the direction of the adjustment. In the worst case, where the bias types augmented each other, the error was on the order of 1.5 dB. Due to the statistical modeling of these bias phenomena, the other parameters of the model, the *SL* values, will obtain estimates in which the bias errors are taken out.

4.4. Accuracy of Results

The GLM model uses the *DifferenceLevel* values resulting from every adjustment performed in the experiment to estimate the set of *SegmentLevel* values. Thus the variance, or uncertainty, associated with the individual adjustments is to some degree evened out. Moreover, the effects of two types of bias errors are reduced, as detailed in the previous section. The accuracy of the loudness estimates based on the model can be expressed either as the standard error of the model parameters, or as the variance in the residual error corresponding to the ε term (eq. 11).

Because the experiment design was balanced, as described in section 2.7, the standard error is the same (within 0.5% variation) for all the *SL* parameters. For the model specified in eq. 11, the standard error of the *SL* parameter estimates is 0.25 dB. In other words, 0.25 dB is the "expected deviation" of the estimated loudness level for each sound segment, if the experiment would be repeated. This corresponds to a 95% confidence interval of ±0.49 dB. The residual error, in a regression model, is commonly specified as the root mean-square-error (RMSE). This model's RMSE is 1.85 dB (which, in our case, might be less relevant than the uncertainty of the *SL* estimates).

The accuracy obtained from the model parameters, e.g. the *SL* values, is implicitly controlled via the degree of redundancy in the experiment design. Up to $n*(n-1)/2$ different matches of the n segment pairs could theoretically be made by each subject, without having to do any repetitions. In the main experiment, the number of matches was much closer to n than to n^2. So, combined with the results from the pilot experiment (section 2.7), it is reasonable to presume that the accuracy could have been improved further by including more of the possible loudness matches. To *predict* this increase in accuracy, however, is non-trivial.

4.5. Individual-SL Models

The models of the previous sections all incorporate a set of *SL* parameters which are 'shared' between all subjects, i.e., all of the adjustments obtained in the experiment are used jointly in the estimate of the *SL* parameters. Alternatively, a model could contain a set of *SL* parameters for each individual subject. This particular GLM model, which would be an extension of eq. 11 containing both individual bias terms and individual *SL* parameters, is specified in eq. 12. In this model, the adjustments (*DL*) of a given subject have no influence on the *SL* estimates of any other subject.

$$DL(i,j) = SL_S(i) - SL_S(j) + sign(AdjLevel) \cdot AdjBias_S + ABbias_S + \varepsilon_S \qquad (12)$$

4.5.1. Within-subject inconsistency

A simple way of measuring the within-subject variability – or inconsistency – of each subject would be to let the subject match each segment pair a number of times, and then estimate the variance of the adjustments. However, in the experimental method described here, no adjustments were 'wasted' on such repetitions – instead adjustments were obtained for a larger number of different pairs.

By modeling the responses of each subject using individual model parameters (eq. 12), the inconsistency of the individual subjects can be studied by *assuming that the adjustments of a perfectly consistent subject would fit the model completely*. The inconsistency can then be quantified by measuring how much a subject's adjustments deviated from the predictions of the subject-specific model, and how accurate the *SL* parameters can be estimated (similar to section 4.4). The first method can be said to measure the inconsistency in the *DL*-domain, while the second one measures inconsistency in the *SL*-domain.

The residual error (corresponding to the variance of the ε_S term), ranges from RMSE = 1.53 dB for the least inconsistent subject, to RMSE = 1.94 dB for the most inconsistent. This range of residual error indicates that there was some difference between the inconsistency of the subjects – in particular one subject had a somewhat higher residual error. In a less homogenous group of subjects, we would expect this difference to be more pronounced. The *SL* estimates, for all except the single most inconsistent subject, have a standard error in the range [0.54, 0.68] dB. Again, this observed accuracy depends on the redundancy afforded in the experiment design.

4.5.2. Between-subject agreement on SL

An investigation of the between-subject variability, or *disagreement*, is possible using the model with subject-specific *SL* parameters. The disagreement is measured as the absolute difference between a given subject's SL_S (eq. 12) and the "common loudness" or shared *SL* (eq. 11). If the disagreement is low then the (individual) SL_S

parameters will tend to be close to the shared *SL*, and vice versa.

The study of mean disagreement of same-pair adjustments, measured by the Gini's pairwise mean difference (section 3.2), differs in two ways from the between-subject variability based on the subject-specific *SL* estimates. Partly, the *SL* approach has the advantage that an *SL* estimate for *every segment* is computed via the GLM model, so the entire set of *SL* parameters can be compared, rather than just the segment pairs repeated by different subjects. And partly, by using the *SL* estimates, both the bias and some of the undesired variance due to the adjustment procedure itself (the experimental error) are factored out via the model. The variance due to subjective differences in the perception of loudness, on the other hand, remains in the SL_S parameters.

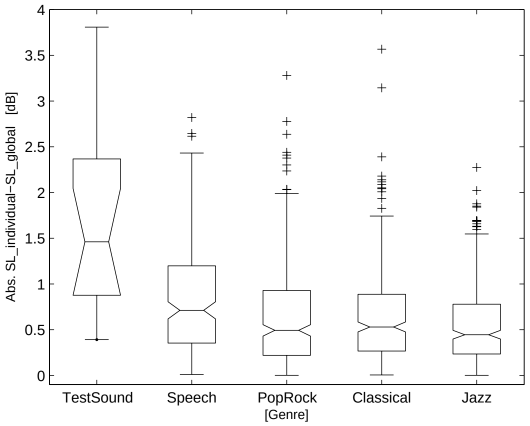


Figure 11. The box-plot for each class of stimulus shows the distribution of absolute difference between the individual SL_S and the common *SL* parameters corresponding to a segment in that class, for all subjects. The 'notch' indicates the CI of the median.

Figure 11 shows the distribution of this absolute difference between *SL* and SL_S, for all subjects combined. The differences are plotted in terms of the class of the segments corresponding to the *SL*-values. Please note that the unequal sample-size for the classes might bias the smaller classes (test-sound and speech) towards a higher disagreement. Jazz music seems to be the most agreed-on class of stimulus, and in general the disagreement for the music genres is below 2 dB. The speech class appears to have higher mean disagreement than the music genres, although the difference is just around 0.2 dB. TestSound is the most disagreed-on class, and section 3.2.2 already identified the 1 kHz tone as the culprit.

The disagreement of all subjects is for all segments below 2.5 dB, excluding a few outliers. Note that these between-subject differences are calculated on the *SL* estimates, i.e., *after* the corrections for the individual bias errors and after the experimental errors have been evened out. The observed differences therefore suggest the existence of a considerable subjective factor in the loudness assessment of music and speech.

4.6. Dependency of Between-subject Variability on Stimuli Properties

The between-subject variability that remains in the loudness assessments, after bias effects and experimental error have been reduced, might be caused by differences in the individual perception of loudness. Such systematic differences between subjects would presumably depend on certain properties of the stimuli. In this section, the GLM model is extended with factors to model or predict the deviation of the individual subjects from the shared *SL* parameters.

Two different methods of predicting the individual deviations from the common *SL* are pursued: (1) Include extra factors which are based on *a priori* information about the stimuli, namely their segment class. (2) Include extra factors which are based on measured signal properties of the stimuli.

4.6.1. Subjective Class Bias

In the process of collecting sound stimuli for the experiment, each segment was classified into one of five different classes or genres (see Table 1). Every segment that was chosen clearly belongs to one and only one of these classes. In particular, the segments within each class were chosen so that they would together constitute a *representative sample* of that class. In other words, the spectral and dynamic properties of the segments vary considerably within a class, while still possessing the sonic properties to assure their membership of the given class.

Loudness is known to depend on spectral-dynamic properties of the stimulus. Models of loudness exist that can accurately predict the subjective loudness of stationary sounds such as tones or noise, based on their spectral properties. However, speech and common music genres constitute a particular type of material because of the listeners' experience and expectations with them. This familiarity may influence their loudness perception.

To model the hypothesis that each subject perceived the stimuli from a certain class louder or softer than other subjects, a *ClassBias* term is added to the model of eq. 11. In eq. 13 the *ClassID(i)* has a value corresponding to the class of segment *i*, with a unique value for each class. The subject-specific *ClassBias* term then adds a constant corresponding to the class of segment *i* and subtracts the constant for the other segment in the pair, *j*. Note that when the two segments belong to the same class, the *ClassBias* terms cancel each other.

$$\begin{aligned} DL(i,j) &= SL(i) - SL(j) \\ &+ sign(AdjLevel) \cdot AdjBias_S + ABbias_S \\ &+ ClassBias_S(ClassID(i)) \\ &- ClassBias_S(ClassID(j)) \end{aligned} \qquad (13)$$

When appending factors to a regression model, i.e. adding extra model parameters, it is appropriate to test whether the model fit is significantly improved. Essentially, this hypothesis is tested with an *F*-test of whether the extended model's R^2 is significantly greater than the R^2 of the simpler model, relative to the larger number of parameters in the extended model [33].

In the case of the GLM models of eq. 11 vs. eq. 13, the test yields: F(28,8389) = 21.26, p < 0.0001, meaning that the model with the ClassBias is significantly better.

By computing the Type III sum of squares (SS), or the partial sum of squares, in the ANCOVA, it is tested whether each factor significantly reduces the residual SS of the model in which all other factors are included [34]. Such *F*-tests, included in Table 6, show that the ClassBias factors are significant.

Source	*DF*	*Type III S.S.*	*M.S.*	*F*	*Pr > F*
ABBias	8	614	76.8	22.29	<.0001
AdjBias	8	2337	292.2	84.78	<.0001
$\text{ClassBias}_{\text{Speech}}$ × Subj	7	670	95.7	27.79	<.0001
$\text{ClassBias}_{\text{PopRock}}$ × Subj	7	409	58.5	16.98	<.0001
$\text{ClassBias}_{\text{Classical}}$ × Subj	7	434	62.0	18.01	<.0001
$\text{ClassBias}_{\text{Jazz}}$ × Subj	7	445	63.6	18.48	<.0001

Table 6. ANCOVA, test of significance for the ClassBias factors.

In Table 6, *ClassBiasN×Subj* indicates a *ClassBias* term for class N, for each of the 8 subjects, yielding 8–1=7 *df* as the *ClassBias* parameters are only specified relative to each other. Due to over-parameterization the *ClasBias* factors for one class must be left out of the model, in order to perform the ANCOVA and parameter estimates. Hence, the $ClassBias_{TestSound}$ is not included in the test displayed in Table 6. As for the actual parameter estimates, nearly all the *ClassBias*-parameters are in the range [–1.5, 1.5] dB. For a given class, the set of *ClassBias*-parameters for all subjects will be centered at 0 dB.

In summary, the deviation of the individual loudness assessments from the 'shared' or common *SL* estimates depend significantly on the class of the stimuli. The magnitude of this individual deviation is up to ±1.5 dB.

4.6.2. Four MPEG-7 Audio Descriptors to Characterize the Stimulus

Each segment belongs to a class: either speech material or test sounds, or a music genre, such as pop/rock or classical. However, a different classification of the segments could be conceived: the classes could depend on certain spectral and dynamical properties – also called acoustical features or signal features – of the segments. Such features would characterize the signal's bandwidth, spectral balance, dynamic range etc. These properties of each segment could be measured by signal processing analyses, and the segments could be classified accordingly. The class dependency investigation, presented in the preceeding section, could then be repeated using the spectral-dynamics based segment classification instead of the *a priori* classes.

In the Audio part of the *MPEG-7* standard a number of Low-Level Descriptors (LLD) are defined [35]. They provide normative methods of characterizing various acoustical aspects of audio signals. The MPEG-7 LLDs are not intended for a specific application, but are meant as low-level features that may be transformed and combined depending on the application.

Four MPEG-7 LLDs were employed to measure four different properties of each stimulus. These four selected LLDs are (based on definitions from [35]):

- **Spectral envelope**
  Describes the short-term power spectrum of the audio waveform as a time series of spectra with a logarithmic frequency axis. The spectrum consists of a series

of coefficients representing power in logarithmically spaced bands between the specified edge frequencies.

- **Spectral flatness**
  Describes the flatness properties of the short-term power spectrum of an audio signal within each of a given number of frequency bands. This descriptor expresses the deviation of the signal's power spectrum over frequency from a flat shape (corresponding to a noise-like or an impulse-like signal). A high deviation from a flat shape may indicate the presence of tonal components.
- **Spectral centroid**
  Describes the center of gravity of the log-frequency power spectrum. The SpectrumCentroid is defined as the power weighted log-frequency centroid.
- **Spectral spread**
  Describes the second moment of the log-frequency power spectrum. Spectrum spread is an economical descriptor of the shape of the power spectrum that indicates whether it is concentrated in the vicinity of its centroid, or else spread out over the spectrum. It allows differentiating between tone-like and noise-like sounds.

Four new meta-features were defined, based on the four MPEG-7 Audio LLDs. These meta-features, listed in Table 7, characterize four different aspects of an audio segment which may influence the difficulty or subjectivity of judging the relative loudness of a segment pair.

Meta-feature name	Based on MPEG-7 LLD	Definition of meta-feature
SpeEnvStd	AudioSpectrum-EnvelopeType	The SpectrumEnvelope is transformed into dB. For each frequency band, the variance over time is calculated. A resolution of 1 octave is used, and the highest and lowest frequency bands are discarded. The SpeEnvStd is then the square-root of the average of the variance of the remaining bands.
SpeFlaMean	AudioSpectrum-FlatnessType	The mean value over time is calculated in each default frequency band in the SpectrumFlatness. The 1% highest and lowest values in each band are excluded (i.e. a trimmed mean). The SpeFlaMean is then the average across the bands.
SpeCenMean	AudioSpectrum-CentroidType	The median over time of the SpectrumCentroid is used as the SpeCenMean.
SpeSprMean	AudioSpectrum-SpreadType	The median over time of the SpectrumSpread is used as the SpeSprMean.

Table 7. Definition of the four new meta-features based on four MPEG-7 Audio LLDs .

In summary, the *SpeEnvStd* measures the dynamics in octave bands, the *SpeFlaMean* measures the noisiness vs. tonalness, the *SpeCenMean* measures the spectral center of gravity, and the *SpeSprMean* measures the variability of that center. Each of the meta-features generates a single scalar value for each sound segment and the frame-based LLDs are averaged over time. For a pair of sound segments (*i, j*) their difference with respect to one of the meta-features can be computed as:

$$SpeEnvStd_{Diff}(i,j) = SpeEnvStd(i) - SpeEnvStd(j) \quad (14)$$

SpeFlaMean$_{Diff}$, *SpeCenMean*$_{Diff}$, *and SpeSprMean*$_{Diff}$ are defined similarly. These four factors measure how different the two segments in a given pair are, in terms of the properties measured by the meta-features.

In the field of Music Information Retrieval (MIR), several studies investigated how the music genres are related to – and can be predicted from – signal features similar to the ones above (e.g., [36, 37, 38]). In such studies, an entire set of features is often used in linear and nonlinear combinations, to achieve the best prediction. In this section, we are *not* trying to model the relation between the music genres and the signal features. Instead, we consider them as two *alternative* ways of characterizing the segments used as stimuli. For each of the MPEG-7 based meta-features (Table 7) we examine its ability to account for the between-subject disagreement in loudness ratings. Whereas a better model fit might be obtained by using the features in combination, we get an overview of their influence by examining them individually.

The four difference-factors (as eq. 14) are added to the model of eq. 11 to form the GLM model in eq. 15. This model supports the hypothesis that each subject perceives loudness somewhat differently than other subjects, depending on differences in certain spectral-dynamic properties of the segment pair. For instance, SCM_S means that a coefficient to the *SpeCenMean*$_{Diff}$ is assigned to each subject, and is estimated along with the other model parameters.

$$\begin{aligned} DL(i,j) = SL(i) - SL(j) \\ &+ sign(AdjLevel) \cdot AdjBias_S + ABbias_S \\ &+ SCM_S \cdot SpeCenMean_{Diff}(i,j) \\ &+ SES_S \cdot SpeEnvStd_{Diff}(i,j) \\ &+ SFM_S \cdot SpeFlaMean_{Diff}(i,j) \\ &+ SSM_S \cdot SpeSprMean_{Diff}(i,j) \end{aligned} \qquad (15)$$

Analogous to the previous section, the significance of each new factor is tested, by computing the Type III sum of squares. Table 8 shows that all four difference-factors are highly significant in predicting the subject-deviation from the 'common' loudness predicted by the *SL* parameters. In particular, the *SpeEnvStd* and the *SpeSprMean* meta-features obtain a high *F* value. This means that subjects tended to disagree relatively more about the loudness of a segment pair in which the two segments had different octave-band dynamics, and/or when the two segments have different degrees of variability of their spectral centroids.

Again, as in the previous section, it is tested whether the extended model has a significantly lower residual error than the simpler model (i.e., without the new factors). And again, this test is positive: F(28,8389) = 33.40, p < 0.0001.

Source	*df*	*Type III S.S.*	*M.S.*	*F*	*Pr > F*
ABBias	8	631	78.9	23.76	<.0001
AdjBias	8	1974	246.7	74.29	<.0001
$SpeCenMean_{Diff}$ × Subj	7	271	38.8	11.70	<.0001
$SpeEnvStd_{Diff}$ × Subj	7	1110	158.6	47.75	<.0001
$SpeFlaMean_{Diff}$ × Subj	7	180	25.7	7.76	<.0001
$SpeSprMean_{Diff}$ × Subj	7	885	126.4	38.08	<.0001

Table 8. Test of significance for bias terms and LLD-based factors.

4.7. Reference Points and Constraints for Model Parameters

In the GLM models presented in the previous sections each observation or data entry has consisted of a *DifferenceLevel* (*DL*) which is modeled in terms of a difference between two *SL* parameters (eq. 2, eq. 5). Therefore, as all the *DL* observations are concerned with *relative SL*, an absolute set of *SL* parameter values cannot be estimated. For example, consider the ANCOVA in Table 3. This GLM model contains a set of 147 *SL* parameters corresponding to the 147 different sound segments. However, the ANCOVA lists only 146 *df* (degrees of freedom) as the *SL* parameters are only specified relative to each other. To obtain a unique estimate of the model parameters, some *fixed point* or *reference point* needs to be added to the data set.

A reference point can be specified as in eq. 16, where refSegmID is the segment which is used as reference or anchor point, and RefPoint is its *SL*-value. Note that the reference point should have a zero coefficient for any bias terms of the model, as it will otherwise be biased.

$$RefPoint = SL(refSegmID) \qquad (16)$$

The 1 kHz pure tone that was included as a stimulus in the experiment can be selected as the reference point. The *loudness level* of a given sound is defined as the SPL of a 1 kHz tone with the same perceived loudness [17]. By setting the *SL* value for the 1 kHz tone segment equal to its measured SPL, all the estimated *SL* values will then by definition correspond to the loudness level, and thus a *phon scale* is obtained. In the main experiment *RefPoint = 70 phon* can be used, because the experiment was SPL-calibrated to this level.

Section 4.5 introduced a GLM model with individual *SL* parameters – the SL_S. It is not obvious, however, how the different sets of SL_S parameters for the different subjects should be aligned with each other. One possibility would be to simply re-use the reference point approach (eq. 16), replicated for the different subjects. This would align the SL_S parameters precisely at the reference point (e.g. 70 phon for the 1 kHz tone). But imagine a situation where two subjects disagreed about the *SL* for the reference sound, but agreed on the other segments. In this case, the single-reference-point approach would lead to a poor alignment of the SL_S parameters.

Instead, the pairwise distance between each SL_S parameter and the corresponding shared *SL* parameter can be minimized. In eq. 17, *SL* represents the 'shared' or 'all-subjects' *SL* parameters, and SL_S represents the individual *SL* parameters for a given subject. The equation shows how this linear constraint can be used as an alternative reference point, by aligning the sum of *SL* parameters so that a unique set of SL_S parameters can be estimated. Specifically, the expression ensures that the RMS difference, between the *SL* and the SL_S parameters for a subject, is minimized.

$$\sum_i (SL(i) - SL_S(i)) = 0 \Leftrightarrow \sum_i SL(i) = \sum_i SL_S(i) \quad (17)$$

A similar situation applies to the *ClassBias* model parameters (section 4.6.1), which are also specified relative to each other. In this case, the following constraint centers the *ClassBias* values of all subjects (for a given class).

$$\sum_S ClassBias_S(ClassID) = 0 \quad (18)$$

Unfortunately, the addition of the reference points or linear constraints changes the variance structure of the model's data. Even though, from a regression point of view, the *SL*-parameters would be estimated correctly, certain statistics of the ANCOVA would depend arbitrarily on the *RefPoint* value. To avoid this problem, all ANCOVA listings and hypothesis tests concerning model parameters that are displayed in the preceding sections are based on the models without any reference points or constraints added.

5. DISCUSSION

This study has been concerned with loudness matching of music and speech material. In this respect, it differs from traditional psychoacoustical studies on loudness perception and just noticeable differences (JND), in which the pair of stimuli to be matched will typically differ *only* with respect to loudness or specific variations of spectral content [17]. Several previous studies on loudness involving broadcast material as stimuli were concerned with loudness *preferences* of listeners [39, 5, 40]. Whereas the loudness matching task used in the present study is probably not independent of subjective preference factors, both the instruction to the subjects and the experiment itself were constructed to suppress it.

The statistical modeling of the responses from the experiment was developed to suppress the experimental error and systematic errors. Two kinds of bias phenomena were given careful attention. In particular, the adjustment bias is well known from psychoacoustic (and other psychophysical) experiments. For example, "*When measuring loudness level of an object sound by the method of adjustment, one type of experiment involves adjusting the object sound until it is as loud as the standard sound. However, a second experiment is needed to remove the bias from the experimental results.*" [17] (p.208). To evade the adjustment bias, loudness matching has been compared to loudness scaling [41], and to adaptive up/down methods [25].

To evaluate their "Loudness Indicator", Jones and Torick performed a loudness matching experiment using the MOA [3]. The stimuli consisted of 2-second segments of processed and unprocessed broadcast material. This experiment used a noise segment as a fixed reference, but matches were performed both in the A/B and B/A order to avoid biased results. The between-subject disagreement was measured using the inter-quartile range (IQR), which is the range of the middle 50% of the values. For the processed broadcast material, the IQR varied between 2.5 and 4.5 dB, and for the unprocessed material, between 1.0 and 3.5 dB.

In september 2000, the WP6P of the ITU-R created a Special Rapporteur Group, the SRG-3, to investigate into "Audio metering characteristics suitable for use in digital sound production" [42, 43]. Specifically, the ITU-R considered "*...that listeners desire audio programmes to be uniform in subjective sound level*", and thus posed the question: "*What audio metering characteristics should be used to provide accurate indication of subjective programme loudness?*". In order to decide on this question, a set of subjective reference data was produced by means of a loudness assessment experiment. In the experiments, a segment consisting of female speech was used as a fixed reference. A total of 49 sound segments were collected to be representative of broadcast programme material. In the collection, 13 of the segments are music, and the rest are primarily speech material with or without various background sounds. Listening experiments have been conducted at 5 different test sites. Certain details of the experiments were presented by the Australian National Acoustic Laboratories and the Australian Broadcasting Corporation [44].

The "Experiment 1" in [21] was a loudness matching experiment performed using the MOA and a sample of broadcast material as stimuli. A 10-second female speech segment was used as a fixed reference. The experimental design and the collection of segments were reused from the ITU SRG-3 study. Each subject matched each stimulus against the reference twice, once with a positive level offset added to the stimulus and once with a negative offset. Thus, the within-subject consistency was assessed by measuring the difference in these same-subject, same-stimuli adjustments. The mean absolute difference was 1.24 dB, averaged across subjects. No adjustments were made in the reverse A/B order, and an adjustment-bias effect was observed, on the order of 0.5-1.0 dB. The between-subject disagreement was assessed by the standard error of the mean adjustment level. This standard error was around 0.6 dB for non-speech segments, and 0.4 dB for speech material. It was concluded that "*a higher level of agreement was found among subjects when matching the loudness of speech-based signals as compared to music or sound effects*." However, the bias caused by using a speech segment as a fixed reference was not considered.

In the MEDUSA project, numerous aspects of multi-channel level alignment were studied. Initially, several test signals with psychoacoustically motivated spectral content were generated with the intent of minimizing the within-subject inconsistency [19]. Based on subsequent experiments, it was discovered that the calibration signal characteristics were not significant, whereas the source distance was the dominating factor [20]. A kind of between-listener disagreement was measured: the standard deviation of matches between channels was found to be in the range 0.4 to 0.6 dB. An adjustment bias was significant, and was suppressed in the model of the responses.

In [23, 22], subjects aligned the loudness of different loudspeakers using noise and music stimuli. The objective was – via this loudness equalization – to eliminate loudness differences as a factor in subsequent tests of subjective quality of loudspeakers. The subjects indirectly adjusted the level of individual loudspeakers by "giving signals" to the experimenter. A between-subject disagreement was noted, but not quantified: "*There remains a substantial residual variance however, due, in addition to normal errors such as inadvertence of the subjects, to the lack of a full consensus among subjects*."

In our experiment, the between-subjects disagreement of same-pair adjustments was generally lowest when matching two segments from the same musical genre. In particular, the disagreement was exceptionally high for pairs in which one of the segments was the 1 kHz pure tone. We found no evidence that using a speech segment as one of the segments in a pair could lead to a lower between-subject variability. This is in contrast to experiments using a speech segment as a fixed reference.

There is no guarantee that other bias phenomena are not affecting the results. In particular, in a larger experiment performed at multiple sites, the acoustics of each site's room and loudspeakers, and variations of instructions to the subjects, might contribute with different types of bias. In general, when a bias phenomenon is not modeled, two situations exist: 1) if the factors that cause a bias have been randomized in the experiment design, they will add noise (extra variance) to the responses. 2) if the factors were not randomized, e.g. because they were not known or could not be controlled in the experiment design, the responses will contain a systematic error. It would be interesting to extend the models presented here, to incorporate site-specific bias terms. By doing so, the potential extra bias factors in a distributed experiment might be quantified and optionally removed.

The test subjects in our experiment were expert listeners enrolled in the same sound recording program. One could therefore expect that a random sample of "ordinary" listeners as subjects would constitute a less homogeneous group, which might lead to both a larger degree of within-subject inconsistency and perhaps also a larger between-subject disagreement than found in our experiment.

In future work, a loudness assessment experiment might be conducted using multi-channel stimuli, for instance, 5.1 format material. In everyday listening situations, mono material is somewhat rare. To further increase the authenticity of the experiment, the influence from the listening room could be included: colorations and reflections of the stimulus, both of which may influence the loudness perception.

From a methodological perspective, it would also be interesting to conduct a systematic comparison of bias phenomena, accuracy, and efficiency aspects, of loudness experiments using different experimental methods: (1) loudness matching, using a rotary knob, as used here, (2) using key-presses to control level, as in many of the recent experiments reported above, (3) using adaptive up/down methods, (4) possibly compared with loudness scaling methods. To our knowledge, no such systematic study exists in connection with the long-term loudness of music and speech material.

6. CONCLUSIONS

A loudness matching experiment was conducted, using the method of adjustment. The relative level of one of the segments in each pair was controlled by the subject using an endless rotary knob. The stimuli consisted of 147 homogeneous sound segments collected as representative samples from 5 classes of sound: 3 genres of music, together with speech and two test sounds. The dynamic range of the experiment was controlled by means of a level normalization using a pseudo-loudness function followed by a stochastic spread of the presentation level.

A pilot experiment allowed a comparison of a method based on matching all stimuli against a single fixed-reference segment with a method in which both segments of a pair were selected among all of the stimuli. The experimental design was constructed with redundancy, in the sense that each subject performed more than one match involving each segment. By affording a certain redundancy and by using a balanced experiment design, the accuracy of the results of the latter method was improved beyond the former method with any segment as fixed-reference.

In the main experiment, when two subjects performed a loudness match of the same pair of segments, their adjustments were typically 2.1 dB apart. Equivalently, the typical standard deviation of same-pair adjustments was 1.7 dB. These figures describe the direct level adjustments which include both subjects' biases and experimental errors. The lowest mean disagreement was observed when segments from the pop-rock or jazz genres were compared to other segments from the same class. When matching speech against speech, the mean disagreement was slightly higher. The highest disagreement was found when matching a 1 kHz tone against music segments.

A statistical model of the responses was developed. Two types of bias were included in the model: one related to the A/B-order of the segments, and another related to the direction of the adjustment level. By modeling the systematic errors caused by these two bias phenomena, the magnitude of the bias errors was estimated, and their effects on the results were removed. Both types of bias were found to be subject-dependent. When considering both bias types together, adjustments would deviate by 1.5 dB from unbiased adjustments.

The *common loudness*, i.e., the best estimate of the loudness level of every stimulus based on *all* level adjustments, was computed via the model. The resulting estimates had a standard error of 0.25 dB. This accuracy was considerably higher than the accuracy of the individual adjustments.

An individual adjustment made by a subject will deviate from the adjustment predicted from the common loudness. This deviation was modeled in terms of three different phenomena: within-subject inconsistency, between-subject disagreement, and bias phenomena.

The within-subject inconsistency was measured as the residual error of subject-specific models. The RMS error was between 1.5 and 1.9 dB, depending on the subject.

The model was extended to account for between-subject disagreement, i.e., systematic deviations from the common loudness. In contrast to the within-subject inconsistency, the between-subject disagreement is the difference in the subjects' perception of loudness. In other words, a different experimental design would presumably detect a similar disagreement. The between-subject disagreement was modeled in two different ways: First, using (*a priori*) class/genre information about the stimuli. Second, using signal properties of the stimuli measured with MPEG-7 Audio Descriptors characterizing the differences in spectral and dynamic properties of sound segment pairs. Both methods could predict with significance some of the subjective deviations from the common loudness.

In summary, the responses of this experiment were modeled by several factors. The most influential factor was the common loudness, i.e., the contribution from the individual loudness levels of the segments. The less influential factors were the adjustment bias, the A/B-order bias, the between-subject disagreement (as predicted by the segment classifications), and the within-subject inconsistency (a random error).

The results from a loudness assessment experiment may be utilized as "subjective reference data" against which algorithmic measures of loudness are evaluated. The desired accuracy for such loudness measures is of the same order of magnitude as the experimental error and bias phenomena associated with the method of adjustment. For this application, it would therefore be desirable to employ an experimental method in combination with a statistical model which would reduce systematic errors and enable a control of accuracy. Such an experimental method and analysis model were presented in this paper. When the within-subject inconsistency and bias errors are reduced as much as practically possible, a certain between-subject disagreement remain. Our experiment indicates that this subjectivity is a minor yet significant ingredient in the perceived loudness of music and speech.

7. ACKNOWLEDGEMENTS

We would like to thank the students in the Sound Recording program at McGill University who participated as subjects in our listening experiments. Also thanks to Gorm Gabrielsen and Jørgen Granfeldt for statistical inspiration, Prof. Wieslaw Woszczyk for sharing insights about loudness, and to Anne-Marie Sylvestre for her invaluable assistance in collecting and editing sound segments.

8. REFERENCES

[1] Moore, B.C.J., Glasberg, B.R. & Stone, M.A. (2003) "Why Are Commercials so Loud? -- Perception and Modeling of the Loudness of Amplitude-Compressed Speech", Journal of the Audio Engineering Society, vol.51:12, pp.1123-1132.

[2] Bauer, B.B. et al. (1967) "A Loudness-Level Monitor for Broadcasting", IEEE Trans.on Audio and Electroacoustics, vol.AU-15:4.

[3] Jones, B.L. & Torick, E.L. (1982) "A New Loudness Indicator for Use in Broadcasting", in Preprint 1878 from the 71st AES Convention, Montreux.

[4] Stone, M.A., Moore, B.C.J. & Glasberg, B.R. (1997) "A real-time DSP-based loudness meter", in Contributions to Psychological Acoustics.

[5] Emmett, J. & Girdwood, C. (1994) "Programme Loudness Metering", in AES UK Managing the Bit Budget Conference, pp.92-98.

[6] Emmett, J. & Emmett, J. (2003) "Audio levels - in the new world of digital systems", EBU Technical Review, vol.2003:January.

[7] ISO (1975) "Acoustics. Method for calculating loudness level. International Standard ISO 532 (1.ed.)", International Organisation for Standardisation.

[8] Moore, B.C.J., Peters, R.W. & Glasberg, B.R. (1996) "A revision of Zwicker's loudness model", Acta Acustica, vol.82, pp.335-345.

[9] Torick, E.L., Allen, R.G. & Bauer, B.B. (1968) "Automatic Control of Loudness Level", IEEE Transactions on Broadcasting, vol.BC-14:4, pp.143-146.

[10] Spikofski, G. & Klar, S. (2004) "Levelling and Loudness - in radio and television broadcasting", EBU Technical Review, vol.2004:Jan.

[11] Klar, S. & Spikofski, G. (2002) "On levelling and loudness problems at television and radio broadcast studios", in AES 112th Convention, Munich.

[12] Lund, T. (2003) "Loudness Control in Digital Broadcast", in Broadcast Asia 2003.

[13] Guildford, J.P. (1954) "Psychometric methods", New York: McGraw-Hill.

[14] Nunnally, J.C. & Bernstein, I.H. (1994) "Psychometric Theory (3.ed.)", New York: McGraw-Hill.

[15] Bradley, R.A. (1984) "Paired comparisons: Some basic procedures and examples", in Handbook of Statistics, Krishnaiah, P.R. & Sen, P.K. (eds.), pp.299-326, Amsterdam: Elsevier.

[16] Gabrielsen, G. (1999) "Paired comparisons and designed experiments", Food Quality and Preference, vol.11, pp.55-61.

[17] Zwicker, E. & Fastl, H. (1999) "Psychacoustics: Facts and Models (2.ed.)", Springer Series in Information Sciences, 22, Berlin: Springer-Verlag.

[18] Poulsen, T. (2002) "Psychoacoustic Measuring Methods", Lecture note no. 3108-e, Lyngby, Denmark: Ørsted - DTU, Acoustic Technology.

[19] Suokuisma, P., Zacharov, N. & Bech, S. (1998) "Multichannel level alignment, Part I: Signals and

methods", in AES 105th Convention, San Francisco.

[20] 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.

[21] Soulodre, G.A., Lavoie, M.C. & Norcross, S.G. (2003) "The Subjective Loudness of Typical Program Material", in 115th Convention of the AES.

[22] 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.

[23] Aarts, R.M. (1991) "Calculation of the loudness of loudspeakers during listening tests", Journal of the Audio Engineering Society, vol.39, pp.27-38.

[24] Cardozo, B.L. (1965) "Adjusting the method of adjustment: SD vs DL", Journal of the Acoustical Society of America, vol.37:5.

[25] Lydolf, M. (1999) "Comparison of eight psychophysical methods for measurement of the threshold of hearing", in The threshold of hearing & contours of equal loudness, pp.18-47, Acoustics Laboratory, Aalborg University.

[26] Verhey, J.L. (1999) "Psychoacoustics of spectro-temporal effects in masking and loudness perception", Oldenburg, Germany: BIS Verlag.

[27] Florentine, M., Buus, S. & Poulsen, T. (1996) "Temporal integration of loudness as a function of level", Journal of the Acoustical Society of America, vol.99:3, pp.1633-1644.

[28] Buus, S., Florentine, M. & Poulsen, T. (1997) "Temporal integration of loudness, loudness discrimination, and the form of the loudness function", Journal of the Acoustical Society of America, vol.101:2, pp.669-680.

[29] IEC (1979) "IEC 60651, Sound level meters", International Electrotechnical Commission.

[30] Howell, D.C. (2002) "Statistical Methods for Psychology" (5. ed.), Duxbury.

[31] Trochim, W.M.K. (2003) "Research Methods Knowledge Base", Internet web site: http://trochim.human.cornell.edu/kb/.

[32] Box, G.E.P., Hunter, W.G.J. & Hunter, S. (1978) "Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building", Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons.

[33] Hand, D.J., Mannila, H. & Smyth, P. (2001) "Principles of Data Mining (Adaptive Computation and Machine Learning)", The MIT Press.

[34] SAS Institute Inc. (1999) "Hypothesis Testing in Proc GLM", chap. 30 in SAS/STAT User's Guide,

[35] MPEG (2001) "Information Technology - Multimedia Content Description Interface - Part 4: Audio (part of MPEG-7)", ISO/IEC CD 15938-4.ISO/IEC JTC 1/SC 29/WG 11,

[36] Allamanche, E. et al. (2001) "Content-based Identification of Audio Material Using MPEG-7 Low Level Description", in Proceedings of the ISMIR-2001.

[37] Tzanetakis, G. & Cook, P. (2002) "Musical genre classification of audio signals", IEEE Trans.Speech and Audio Processing, vol.10:5.

[38] Burred, J.J. & Lerch, A. (2003) "A hierarchical approach to automatic musical genre classification", in Proc. of the 6th Int. Conf. on Digital Audio Effects (DAFx-03).

[39] Belger, E. (1969) "The Loudness Balance of Audio Broadcast Programs", Journal of the Audio Engineering Society, vol.17:3, pp.282-285.

[40] Riedmiller, J.C., Lyman, S. & Robinson, C. (2003) "Intelligent Program Loudness Measurement and Control: What Satisfies Listeners?", in Proc. AES 115th Conv., New York.

[41] Appell, J.-E. (2002) "Loudness models for rehabilitative audiology", Doctorate thesis, University of Oldenburg.

[42] ITU-R (2002) "SRG-3 Status Report (2), September 2002", Document 6P/145-E,

[43] Yahoo Groups (2003) "ITU-R reflector for the SRG3 at Yahoo Groups", Internet web site: http://groups.yahoo.com/group/srg3list/.

[44] ITU-R (2003) "Subjective assessment of loudness characteristics (Australia)", Document 6P/16-E,

[45] Gerber, S.E. & Milner, P. (1971) "The Transitivity of Loudness Level", Journal of the Audio Engineering Society, vol.19:8, pp.656-659.

APPENDIX

Transitivity of Loudness

Gerber and Milner [45] studied the transitivity of loudness level for 5 octave bands of pink noise between 125 and 8000 Hz and presented at a level of 70 phons. In the paired comparisons experiment, the noise stimuli were rated against each other, over a range of 21 dB. For each pair, the task of the listener was to judge if the test sound was louder or softer than the reference. The authors found that "*any sound equally loud to any given reference is equally loud to any other sound equally loud to the reference.*", and concluded that transitivity was a property of loudness level.

In order to use a linear model of the *DifferenceLevels* (*DL*) from our experiment, we needed to test whether the transitivity of loudness level would hold for the loudness assessments performed using real-world sound stimuli. In other words, the hypothesis that the indirect loudness assessment, *DL(i,k)* + *DL(k,j)*, is not systematically different from the direct assessment, *DL(i,j)*. In the pilot experiment, all the music and speech segments (i.e. without the two test sounds) were matched once with each other by each subject. The transitivity test was based on the full set of *DifferenceLevel* values for these segments.

The hypothesis was tested by calculating the difference between the DifferenceLevel resulting from an adjustment for a given segment pair, *DL*(*i,j*), and the "indirect" DifferenceLevel, *IDL*(*i,j*), for the same pair. Each *IDL*(*i,j*) corresponded to certain pairs of ratings *DL*(*i,k*) and *DL*(*k,j*), where segment *k* was neither *i* nor *j*. The *IDL* was calculated as the average over these *DL* pairs (eq 19).

$$IDL(i,j) = \frac{1}{nSegments-2} \sum_{k \notin \{i,j\}} (DL(i,k) + DL(k,j)) \qquad (19)$$

Let d(*i,j*) be defined as the difference between the direct and indirect DifferenceLevel for the segment pair (*i,j*) (eq. 20). This difference was calculated for each subject individually, such that a given *DL* was only compared to the *IDL* based on ratings of the *same* subject. Thus, between-subject variability was not an issue here. Whenever *DL*(*i,k*) was needed in eq. 19, but *DL*(*k,i*) was available as an adjustment by that subject, *–DL*(*k,i*) was used in its place; and likewise for *DL*(*k,j*).

$$d(i,j) = DL(i,j) - IDL(i,j) \qquad (20)$$

Based on the sample of n=604 *d*-values, the transitivity was tested by hypothesizing that *d* has a mean value which is significantly different from 0. Hence the null hypothesis was that $mean(d) = 0$, which implied that no systematical difference between the direct and the indirect relative loudness levels existed. A histogram indicates that the *d* variable was normally distributed with mean around 0, so the hypothesis could be tested using a one sample t-test. The t-test gave t = 1.206 (p = 0.228), meaning that at the significance level, *alpha* = 0.05, the null hypothesis could *not* be rejected. Therefore, based on the data from the pilot, there is not reason to believe that the direct and the indirect loudness levels were systematically different. The transitivity also asserts that the *DL* lies on an *interval scale*.

Now, statistical significance always depends on the size of the data sample. With a somewhat larger number of observations it is possible that an observed difference would indeed have been significant. Therefore the effect size can be considered: The best estimate of the difference between the *DL* and the *IDL* is *mean(d)* = 0.089 dB, which must be considered quite small.

Variations of Response Time

Although all our subjects were expert listeners, they were not trained specifically for the task of the experiment prior to their participation in the experiments. Based on the responses from the main experiment, it was examined whether the response time of the subjects would significantly increase or decrease over the course of the experiment. It would have been relevant and important for this investigation also to consider the development of the inconsistency and/or disagreement of the subjects, over time. However, this is difficult to measure because no adjustments were repeated by the same subject.

Out of the 8 subjects in the main experiment, 4 also participated in the pilot experiment. Therefore, two hypotheses were tested: 1) the subjects would become faster during the main experiment; i.e., *ResponseTime* decreases as *RatingNumber* increases, and 2) the response time would decrease less for the 4 pilot subjects than for the 4 non-pilot subjects (as the former group had more prior practice in this particular task when doing the main experiment). The *ResponseTime* measures in seconds the period that the subject uses to determine and submit a rating, and *RatingNumber* is the number of ratings submitted by the subject. A Q-Q plot indicated that the distribution of *ResponseTime* is nearly log-normal, hence the log transformed variable is used here.

The two hypotheses were tested using an analysis of variance (ANOVA), with the log(*ResponseTime*) as the dependent variable, and *PilotSubj* as a 'dummy' variable with a value of 1 for subjects who participated in the pilot and 0 for the others. In the ANOVA both main factors were significant: *RatingNum*, F(1,8581) = 110, $p<0.0001$, and *PilotSubj*, F(1,8581) = 1313, $p<0.0001$. Hence, we accept both hypotheses stated above: The subjects *did* become faster during the main experiment, and this effect *was* less pronounced for the pilot subjects than for the non-pilot subjects.

Response Time and Same-pair Disagreement

If a high between-subject disagreement was in part due to an indecisiveness of the subjects, this would be revealed by a relationship between the GMD_4 measure (see section 3.2) and the response time of the corresponding adjustment, or between the GMD_4 and the number of A/B-comparisons used during the adjustment. This relation was examined by introducing the *mspResponseTime* variable which is the mean same-pair *ResponseTime*, i.e. one *mspResponseTime* value corresponds to one group of *DifferenceLevel*s that are averaged in the GMD_4 measure (eq. 1). Similarly, *mspNumAB* is the average number of A/B-comparisons used in each group of same-pair ratings. The correlation between the GMD_4 and the corresponding *mspResponseTime* or *mspNumAB* was calculated, across the segment-pairs included in the experiment. The rank-correlation (Spearman's rho, r_S) was used because the involved variables have different scales and distributions.

The rank-correlation between the GMD_4 and the *mspResponseTime* was, r_S = 0.07 (p=0.0015), and between GMD_4 and *mspNumAB*, r_S = 0.10 ($p<0.0001$). These values showed that a small but significant positive correlation exists between the GMD_4 and both performance variables, for the groups of same-pair ratings. In other words, there was a tendency for the GMD_4 to be larger when the *ResponseTime* is longer, and/or when more A/B comparisons were used. However, the correlation magnitudes are quite small, indicating that the relation is not very strong. Therefore these two performance variables, *ResponseTime* and *NumAB*, were not used to model the responses.