Audio Engineering Society

Convention Paper 8331

Presented at the 130th Convention
2011 May 13–16 London, UK

Perceived Level of Late Reverberation in Speech and Music

Jouni Paulus[1], Christian Uhle[1], and Jürgen Herre[2,1]

[1] *Fraunhofer Institute for Integrated Circuits IIS, Am Wolfsmantel 33, 91058 Erlangen, Germany*

[2] *International Audio Laboratories Erlangen, Am Wolfsmantel 33, 91058 Erlangen, Germany*

Correspondence should be addressed to Jouni Paulus (`jouni.paulus@iis.fraunhofer.de`)

ABSTRACT

This paper presents experimental investigations on the perceived level of running reverberation in various types of monophonic audio signals. The design and results of three listening tests are discussed. The tests focus on the influence of the input material, the direct-to-reverberation ratio (mixing level), and the reverberation time using artificially generated impulse responses for simulating the late reverberation. Furthermore, a comparison between mono and stereo reverberation is conducted.

It can be observed that with equal mixing levels, the input material and the shape of the reverberation tail have a prominent effect on the perceived level. The results suggest that mono and stereo reverberation with identical reverberation times and mixing ratios are perceived as having equal level regardless of the material.

1. INTRODUCTION

Reverberation is part of practically all natural acoustic environments and plays an important role for the human perception of sound providing valuable information to the listener about the environment in which the sound sources are located. In the area of music production, be it by capturing live sound performances or by mixing individually recorded instruments, the characteristics of the reverberant sound components determine for the perceived subjective quality of the final sound or music recordings to a considerable extent. Specifically, the control of the level and the characteristics of the reverberant components in a mix is an art which tonmeisters and sound engineers are trained to master. In order to enhance our understanding about the way human listeners perceive reverberation, this paper presents some investigations into determining the level of reverberation, as it is perceived by listeners for usual types music and speech.

The perceived level of reverberation depends considerably on the source signal and the shape of the reverberation decay, as was shown by Gardner and Griesinger [3]. In their experiments, the listeners adjusted the level of reverberation in a test signal to perceptually match the reference signal reverberation level. Their main findings included differences in the perceived level between speech and sustained musical signals, an increase in reverberation level with pre-delay, and the almost equal level of mono and stereo reverberations.

The experiments presented in this paper verify and extend these earlier results with more diverse material and a larger number of listeners. In addition, the inter- and intra-listener differences in the perception are addressed. Instead of a level-matching setup used in the earlier study, we propose assessing the perceived amount of reverberation on an absolute scale. The use of a single rating value allows to compare the effects of various signal-related aspects on the perception. The main properties addressed here are the reverberation time (T_{60}), direct-to-reverberation mixing ratio ($d2r$), mono vs. stereo reverberation on monophonic source signal, and the effect of the source signal itself.

Three listening tests were conducted, each having 12–14 listeners, totaling into 25 participants. All the participants had earlier experience on listening tests, but none of them could be considered as experts in the current task. One of the tests was repeated with the same set of participants to allow assessing the intra-listener rating consistency. A second test has partly overlapping stimuli and a completely independent set of participants, allowing a view to inter-listener differences. With this information it is possible to draw some conclusions of the reliability of the obtained subjective data.

The rest of this paper is organized as follows: Section 2 details the conducted listening tests, the test methodology, generation of test stimuli, and the focus points of each of the three tests. Section 3 presents the obtained results starting with a description of the analysis methods applied to the results, then the results of the individual tests, and, finally, analysis of the results. Section 4 discusses the results obtained. Finally, Section 5 ends the paper with conclusions and topics for the future work.

2. LISTENING TESTS

In the following, the three listening tests are presented starting with describing the test stimuli, i.e., the source signals and the reverberation conditions, then highlighting the focus points of each test, and finally describing the practical aspects of running the tests.

	Label	Description
Tests 1 & 2	hardrock	commercial hard rock, dryish
	guitar	solo acoustic guitar, anechoic
	trumpet	solo trumpet, anechoic
	speech	female and male speech, anechoic
	symphony	symphony orchestra, anechoic
	pop	amateur pop/rock, dry
Test 3	accents	classical, anechoic
	cello	solo cello, anechoic
	electro	electronic loop, dry
	electronic	commercial electronic, dryish
	femalt	commercial alternative rock, dryish
	femopera	opera, anechoic
	fempop	commercial pop, dryish
	femspeech	female speech, anechoic
	funk	commercial funk, dryish
	hardrock2	commercial hard rock, dryish
	malepop	commercial pop, dryish
	malespeech	male speech, anechoic
	metal	commercial heavy metal, dryish
	rock	commercial classic rock, dryish
	trumpet	solo trumpet, anechoic

Table 1: The source signals used in the tests.

2.1. Source Signals

Because it was expected from earlier results [3] that the source material has a prominent effect on the perception of reverberation, the test items were selected to represent various signal classes: speech, individual instruments, and music from various genres ranging from classical opera to heavy metal. A majority of the items originated from anechoic recordings, but some commercial recordings with a moderate amount reverberation were included to increase the variety. A more detailed description of the items is provided in Table 1. All test items were produced to be monophonic either by averaging the channels or by using only the left channel, whichever produced aesthetically more pleasing results. The length of the signals was restricted to approximately 4 seconds.

2.2. Reverberation Conditions

The signals presented to the test participants were generated by mixing the source signal with an artificial reverberation signal in the desired mixing ratio, as illustrated in Fig. 1. The reverberation signals were created using artificial impulse responses simulating the diffuse late reverberation with exponentially decaying white noise after the model proposed by Moorer [12] and later for-

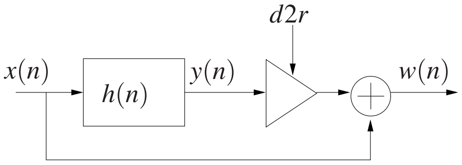

Figure 1: The "wet" signal $w(n)$ is created by mixing the dry signal $x(n)$ with the desired ratio $d2r$ with the reverberation $y(n)$ which results from a convolution between the dry signal and the impulse response $h(n)$.

malized by Polack [13]. In this work, we did not take the early reflections into account. Even though this may cause the reverberation to sound somewhat unnatural, the model has proven to produce a relatively good sounding result with a simple parameterization.

2.2.1. Impulse Responses

The reverberation signal $y(n)$ is a result of convolution between the original signal $x(n)$ and the impulse response $h(n)$ by

$$y(n) = \sum_{i=-\infty}^{\infty} x(n-i)h(i). \tag{1}$$

The impulse responses are generated using the model

$$h(n) = U(n-d)N(n)\mathrm{e}^{-\tau(f)(n-d)/F_s}, \tag{2}$$

where $U(n)$ is the Heaviside step function, d is the pre-delay length in samples, $N(n)$ is a zero-mean Gaussian distributed random noise signal, $\tau(f)$ is a frequency-dependent decay rate, and F_s is the sampling rate. The pre-delay parameter d shifts the entire impulse response in time causing a delay between the direct sound and the reverberation, which is related to the size of the enclosing space and the distance between the sound source and the receiver. Fig. 2 shows an example impulse response generated with this model.

The frequency-dependent decay rate is determined from the desired frequency-dependent reverberation time $T_{60}(f)$ with

$$\tau(f) = \frac{3\log 10}{T_{60}(f)}. \tag{3}$$

For a simplified model, equal reverberation time for all frequencies could be assumed. In reality, however, the high frequencies attenuate faster than the low [16]. Here, the decrease in the reverberation time is estimated to be exponentially decaying with respect to the frequency with

$$T_{60}(f) = T_{60}e^{\lambda f}, \tag{4}$$

where T_{60} is the reference reverberation time and λ is the decay rate parameter. The value of λ used in the experiments was obtained from a fit to the data from [11]. For practical reasons, the frequency range was divided into eight octave bands, and each band had the reverberation time corresponding to its center frequency. A time-frequency representation of an impulse response generated with this frequency-dependent reverberation time is illustrated in Fig. 3.

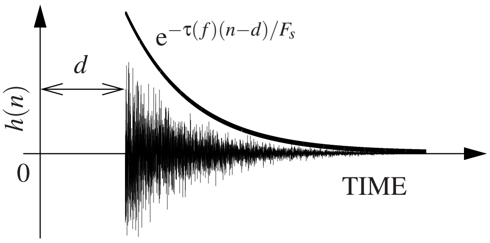

Figure 2: An artificial reverberation impulse response with pre-delay d. The exponential temporal decay envelope applied on the driving noise is illustrated by the solid line.

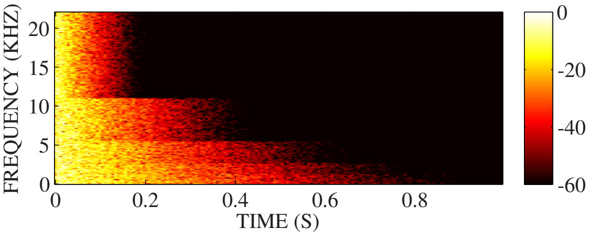

Figure 3: A time-frequency representation of an impulse response generated with the model of Eq. (2). The color encodes the intensity in dB. The reverberation time of the higher octave bands is shorted with the relation of Eq. (4).

2.2.2. Stereo Impulse Responses

For stereo reverberation, separate impulse responses were generated for both channels using the model of Eq. (2) with equal parameters, but using different driving noise signals $N(n)$. The noise signals were generated to have a desired amount of frequency-dependent inter-channel coherence (ICC) [2, p. 238]. The frequency-

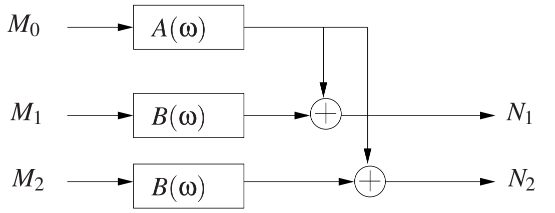

Figure 4: Creating two noise signals N_1 and N_2 with frequency-dependent ICC from three uncorrelated input noise signals M_0, M_1, and M_2 using two filters $A(\omega)$ and $B(\omega)$.

dependent ICC is set to a value of 0.6 from 0 Hz to 350 Hz, with a linear decay to 0 from range 350 Hz to 3000 Hz, and finally 0 above 3000 Hz [15, p. 52]. The desired ICC value for the signals was obtained by using the method of noise mixing [10], [2, p. 243]. In this method, the coherence between two uncorrelated noise signals is increased by adding a third, uncorrelated signal to both signals with a desired level. Making this level frequency-dependent, as illustrated in Fig. 4, the resulting signals will have frequency-dependent ICC. The frequency responses $A(\omega)$ and $B(\omega)$ of two filters used for the level adjustment are related to the desired ICC value $k(\omega)$ by

$$|A(\omega)| = \sqrt{|k(\omega)|}, \tag{5}$$

and

$$|B(\omega)| = \sqrt{1-|k(\omega)|}. \tag{6}$$

The reverberation signal $y(n)$ was mixed with the original signal $x(n)$ by estimating the average loudness of both of them with the method from ITU-R BS.1770 [8] and applying an appropriate scaling to the reverberation component to yield the desired direct-to-reverberation ratio. Finally, the resulting "wet" signals were normalized to have equal average loudness according to ITU-R BS.1770.

The reverberation conditions employed in the first two of the three listening tests are provided in Table 2. The conditions used in the third test are not listed explicitly because there were quite many of them, and the results for the individual conditions are not discussed.

2.3. Listening Test Details

Three listening tests were conducted, each having different objectives. They will be described in the following.

Number	T_{60} (s)	$d2r$ (dB)	ICC
1	1.0	3.0	var
2	1.6	3.0	var
3	2.4	3.0	var
4	1.0	7.5	var
5	1.6	7.5	var
6	2.4	7.5	var
7	1.0	12.0	var
8	1.6	12.0	var
9	2.4	12.0	var
10	1.0	12.0	1
11	1.6	7.5	1
12	2.4	3.0	1

Table 2: Applied reverberation conditions in Test 1 and Test 2. The given T_{60} corresponds to the reverberation time of the lowest band and is lower at higher frequencies as per Eq. (4). The ICC column describes if the reverberation is stereo ("var") or mono ("1"). Pre-delay is 0 ms in all these reverberations.

2.3.1. Test 1

The first test focused on the effect of varying the reverberation time and mixing ratio on various source signals, and on the listener consistency. A total of six source items, as shown in Table 1, and nine stereo reverberation conditions without pre-delay were used (Numbers 1–9 in Table 2).

2.3.2. Test 2

The second test compared mono and stereo reverberations. A subset of three reverberation time and mixing ratio combinations (Numbers 3, 5, and 7 in Table 2) from the first test were selected and the corresponding mono conditions were rendered (Numbers 10, 11, and 12). All the test items were the same as in the first test.

2.3.3. Test 3

The third test focused on inter-listener consistency under more diverse probe signals. The test included 15 items and a total of 36 conditions, but a subset of only 4 conditions for each item was selected pseudo-randomly. The test items cover a broad range of signal classes, see lower part of Table 1. Most of the signals of the polyphonic non-classical music class originated from commercial recordings. Even though they were not anechoic, most of them were found to be quite dry sounding and suitable for the task. The reverberation conditions sampled four parameters: reverberation time, mixing ratio, pre-delay, and ICC. The two first parameters sampled the same three-by-three space as in the first test. As a new parameter, a pre-delay of 50 ms was tested in addition to the 0 ms used earlier. From all of these 18 combinations, both mono and stereo versions were created. Because of the large number of reverberation combinations and original items, it was decided to subsample the conditions such that for each test item, all four pre-delay and ICC combinations were present, but the reverberation times and mixing ratios were selected randomly.

2.4. Test Scheme

The main interest in the tests was to obtain information of the *perceived level of reverberation*, not about more specific aspects such as envelopment, spaciousness, or engagement addressed in some earlier studies. The listeners were asked to rate the amount of reverberation they perceive in the test signals on a scale between 0 and 100 (larger values denoting higher perceived level of reverberation) while ignoring the specific aspects of the reverberation and focusing only on the overall percept.

2.4.1. Anchor Signals

To obtain ratings that can be compared between listeners, direct background anchoring with two signals was used. The anchor signals were presented to the user as additional information and they were not included in the signals to be rated. The anchors were designed to define ratings close to the ends of the scale: a quite dry signal defined a rating value of 10 points, whereas a quite reverberant defined a rating value of 90 points. This anchor positioning allowed overshoot at both ends in order to reduce the amount of potential saturation. The listeners were instructed to utilize the entire scale, and in case the scale was not enough for a certain test signal, indicate that by a textual comment. A similar test design has been used earlier by George et al. [4] for evaluating the envelopment of various signals.

All three tests used the same anchors, and the two anchor signals were based on the same signal, only the amount of reverberation was changed. They represented the signal class of modern instrumental music. One may argue that the anchors should represent all signal classes present in the test for a more reliable comparison between them and the test signals. This approach, however, was rejected after initial experiments for two reasons. First of all, a compound signal becomes relatively long in order to contain excerpts of multiple signal classes. Secondly, and more importantly, all the different component signals should have a perceptually equal amount of reverberation for the anchors to be consistent. This would require a careful adjustment of the reverberation characteristics for each component signal by means of additional listening tests.

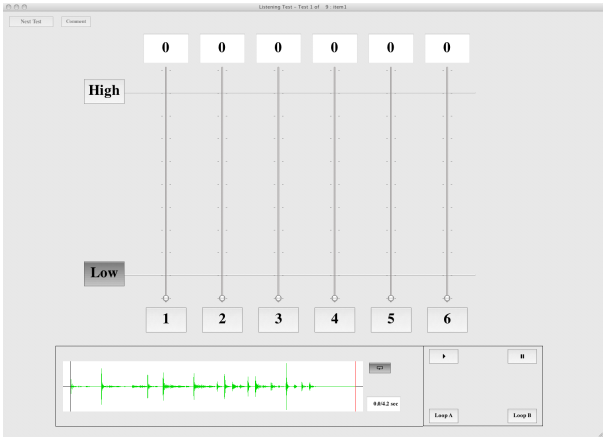

Figure 5: The test interface presented to the participants. The buttons with labels “High” and “Low” denote the anchor signals, while the test signals are denoted with numbers starting from 1. The playback controls are provided in the lower right corner.

2.4.2. User Interface

Each test consisted of multiple graphical user interface (GUI) screens, each of which presented various test signals for the listener to rate. The rating could have been done one test signal at a time, but the setup used was preferred for convenience. Each screen contained stimuli from different signals classes and reverberation conditions. The test signals on each screen were selected pseudo-randomly so that none of the source signals or reverberation conditions were repeated on a screen.

An example of the graphical user interface presented to the participants can be seen in Fig. 5. The two anchor signals were provided by the buttons “High” and “Low” with horizontal lines across the screen providing anchor positions that are easy to locate. The test signals were selected by the numbered buttons, and the connected slider provided a means for entering the associated ratings. The sliders had tick marks on every 10 points for easier visual localization of the rating. The slider values were displayed numerically in the boxes above the sliders.

The test interface was running on a desktop computer located in a separate listening room. The audio was played back with the “Stax SR Lambda Pro (new design)” headphones driven by the “Stax SRM-0061 II” headphone amplifier. The listeners were instructed not to adjust the listening volume during the test.

In the test situation, the participants were first presented with written instructions to the test and a verbal repetition of the main points. Because most of the participants had earlier experience from MUSHRA [7] tests for audio coding quality assessment, the differences between this test setup and MUSHRA were clarified verbally. After the initial instructions the listeners were presented with a familiarization test consisting of only two screens, both with nine signals. The signals were selected randomly from the pool of all test signals and the participants were instructed to rate them. This was done to familiarize the participants with the original signals, with the different reverberation conditions, and with the test interface.

3. RESULTS

This section describes the results of the listening tests, and discusses the listener reliability and the effect of the source material and reverberation conditions to the ratings.

3.1. Evaluation Measures for Listener Reliability

For the obtained subjective data to be reliable, the listeners should agree on the perceived level of reverberation and they should provide similar ratings for repeated stimuli. Given two sets of ratings R_1 and R_2 (from a single listener in a repeated trial, or an aggregate of multiple listeners) four evaluation measures are calculated: root-mean-squared-error (RMSE), mean absolute error (MAE), Kendall tau rank correlation coefficient τ, and Pearson's correlation coefficient r. RMSE and MAE are calculated with

$$\mathrm{L}_p(R_1,R_2) = \left(\frac{1}{K} \sum_{i=1}^{K} |R_1(i) - R_2(i)|^p \right)^{1/p}, \qquad (7)$$

where K is the number of ratings, and the parameter p is 1 for MAE and 2 for RMSE. Kendall tau correlation coefficient [9] compares the orderings resulting from the provided ratings while ignoring the magnitudes. The basic form of the measure can be obtained with

$$\tau(R_1,R_2) = \frac{\sum_{i=1}^{K} \sum_{j=i+1}^{K} c\left(R_1(i),R_1(j),R_2(i),R_2(j)\right)}{\frac{1}{2}K(K-1)}, \qquad (8)$$

where c is an indicator describing the similarity of the pairs by

$$c\left(R_1(i),R_1(j),R_2(i),R_2(j)\right) = \qquad (9)$$
$$\begin{cases} 1, \text{ if } \operatorname{sign}(R_1(i)-R_1(j)) = \operatorname{sign}(R_2(i)-R_2(j)) \\ -1, \text{ otherwise.} \end{cases}$$

The Kendall's tau receives values in the range $[-1,1]$ as with normal correlation coefficient: -1 when the resulting order is reversed, 0 when the orderings are independent, and 1 when they are identical.

3.2. Test 1

The test was taken by 13 participants with the median age of 30.5 a, with the minimum of 25 a and maximum of 51 a. All participants repeated the test after 1–4 weeks, and these data were subsequently used for evaluating the consistency of the individual listener responses.

The main results from the test are provided in Fig. 6; each of the test items are in a separate panel. The mean rating for each reverberation condition over all listeners is denoted by a horizontal line with the 95% confidence interval surrounding it. The conditions are ordered so that the first three correspond to mixing ratio of 3 dB, the next three 7.5 dB, and the last three 12 dB. Within each group of three, the reverberation time is increasing from left to right.

The results show that with equal mixing ratio (the groups of three) increased reverberation time increases the level perceived considerably. The absolute increase is reduced with higher direct-to-reverberation ratios, and the difference appears to exhibit a non-linear behavior with higher levels of direct sound.

Making a simplified assumption that the change in the level perceived depends only linearly on the two physical parameters and that is same with every signal class, a simple parameterization can be obtained for difference in the perceived level Δ_{d2r} based on the the difference in the reverberation time $\Delta_{T_{60}}$ and the difference in the mixing ratio Δ_{d2r} with

$$\Delta_R = 19.2\Delta_{T_{60}} - 2.35\Delta_{d2r}. \qquad (10)$$

It should be noted that this is only a very rough estimate and quite likely will not generalize. However, Gardner and Griesinger [3] reported that reducing T_{60} from 2.0 s to 0.5 s, the level had to be increased by 13 dB to produce equal perception. The parameterization of Eq. (10) suggests a change of 12.3 dB in the mixing ratio given a similar change in the reverberation time.

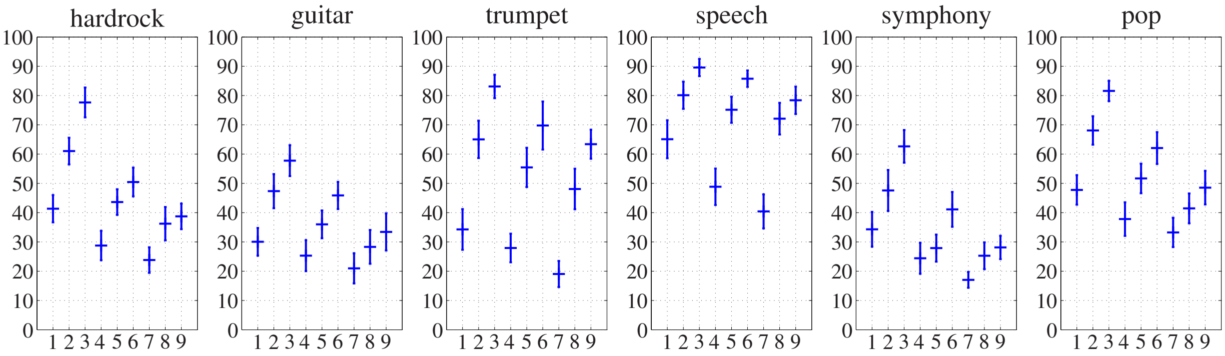

Figure 6: Test 1 Mean ratings (y-axis) for each of the six original items for each reverberation conditions (x-axis). The mean value is denoted with the horizontal line and the surrounding bars denote the range of 95% confidence interval of the rating. See Table 2 for a description of the conditions.

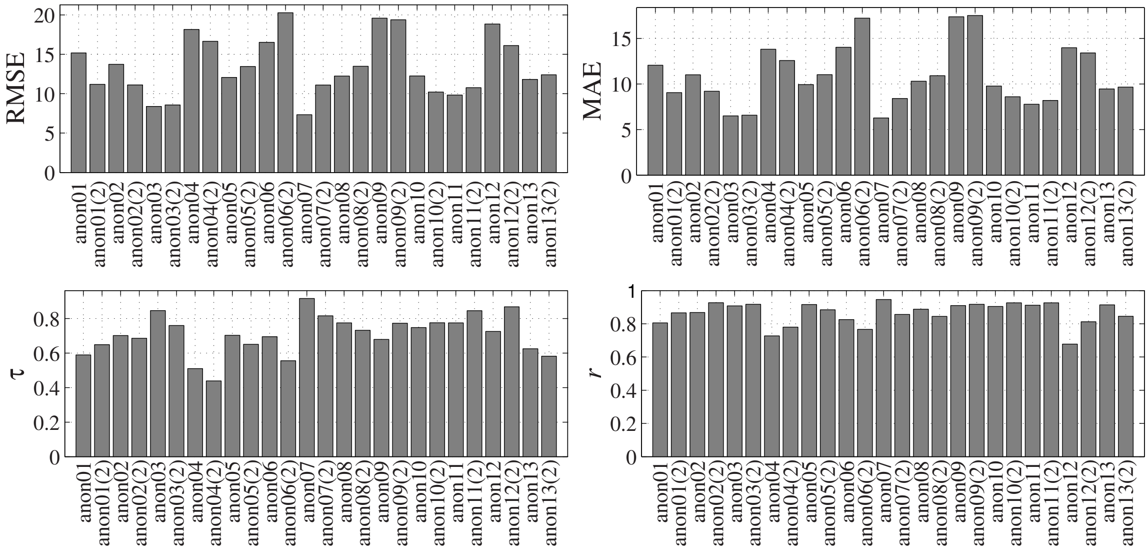

Figure 7: Test 1 Inter-listener evaluation results. For each participation, the evaluation measure is calculated between the participation and the mean of all other participations. Top left panel: RMSE, the mean value is 13.5. Top right panel: MAE, the mean value is 11.0. Bottom left panel: Kendall's tau, the mean value is 0.71. Bottom right panel: correlation coefficient, the mean value is 0.86.

The listener reliability analysis results are shown in Figs. 7 and 8. The listener identities have been anonymized, and the extension "(2)" denotes the repeat of the test. Fig. 7 illustrates the result for each listener session separately using the mean of all other listeners as the reference. It can be seen that there are quite large differences between listeners. The intra-listener results in Fig. 8 show that the differences between repeats of an individual are on average smaller than the differences from the mean rating of all listeners. However, all in all, the differences between inter- and intra-listener results are relatively small.

3.3. Test 2

The test was taken by 12 participants with similar background as in the first test, but none of the listeners took part in the first test. The median age of the participants was 30.5 a, with the minimum 24 a and maximum 37 a.

The main question in the second test was whether the

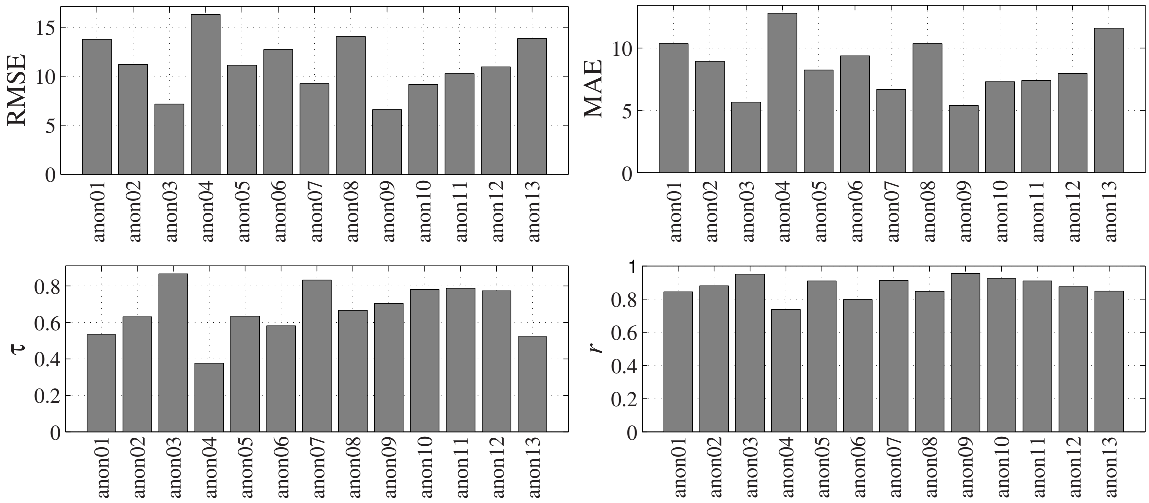

Figure 8: Test 1 Intra-listener consistency evaluation results. Top left panel: RMSE, the mean value is 11.3. Top right panel: MAE, the mean value is 8.6. Bottom left panel: Kendall's tau, the mean value is 0.67. Bottom right panel: correlation coefficient, the mean value is 0.88.

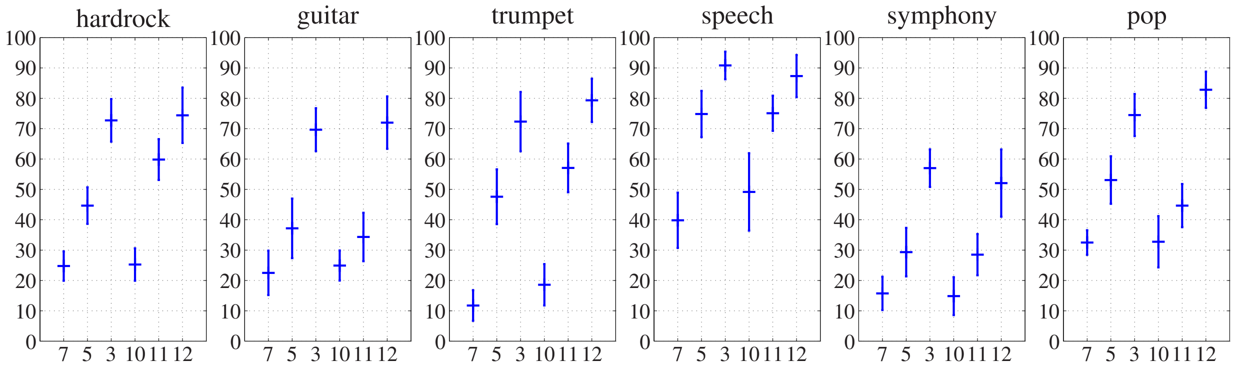

Figure 9: Test 2 Mean ratings (y-axis) for each of the six original items for each reverberation conditions (x-axis). The mean value is denoted with the horizontal line and the surrounding bars denote the range of 95% confidence interval of the rating. Black points are individual listener ratings. See Table 2 for a description of the conditions.

amount of incoherence between the the two channels significantly affects the perceived amount of reverberation. The results, illustrated in Fig. 9, suggest that it does not. In the figure, the three leftmost ratings are for stereo reverberations, while the other three are the same conditions with monaural reverberation. Even though the monaural reverberation ratings mostly exhibit slightly larger values and larger variances, the differences are not statistically significant. Based on Welch's two-tailed t-test with 5% significance level, it is not possible to discard the hypothesis that the stereo and mono reverberation ratings have the same mean value except for one item+reverberation -combination ("hardrock" and reverberation indices 5 and 11 corresponding to approximately mid-level reverberation).

Performing similar statistical analysis between the identical items in the first and the second tests, i.e., only on the stereo conditions, suggests that the mean ratings are equal in both tests, except for the high-reverberation condition of signal "guitar". This in turn suggests that the results of the two sets can be pooled quite reliably, i.e., the monophonic reverberation results are represented on

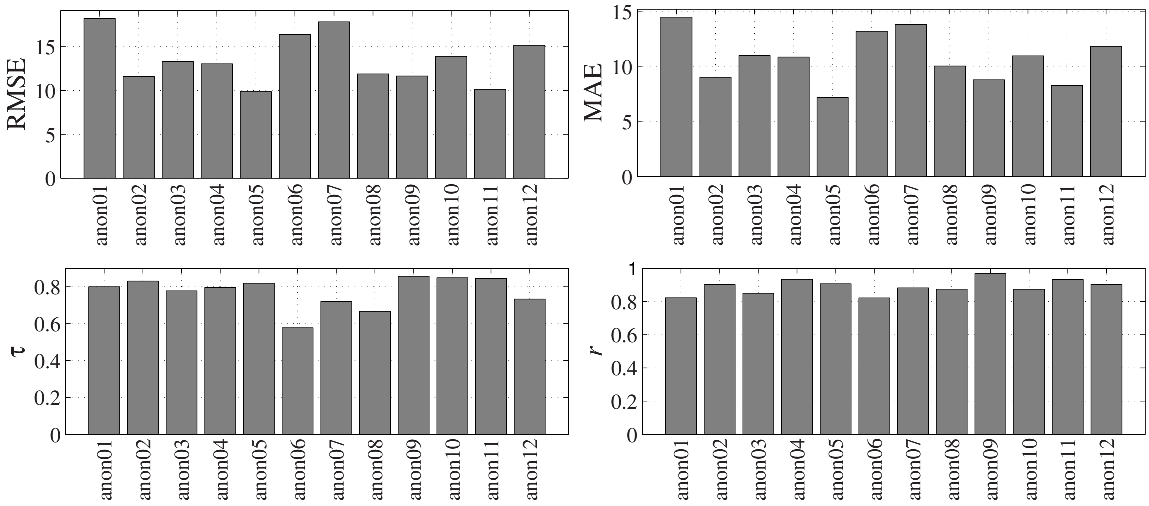

Figure 10: Test 2 Inter-listener evaluation results. For each participation, the evaluation measure is calculated between the participation and the mean of all other participations. Top left panel: RMSE, the mean value is 13.6. Top right panel: MAE, the mean value is 10.9. Bottom left panel: Kendall's tau, the mean value is 0.77. Bottom right panel: correlation coefficient, the mean value is 0.89.

a scale comparable to the scale for the results from the first test. Thus, it can be assumed that the invariance in the rating due to reverberation incoherence applies also in the case of the other conditions.

The inter-listener performance analysis results seen in Fig. 10 resemble the results from the first test. The RMSE and MAE measures between individual listeners and the aggregate ratings are comparable with the first test, and the slight increase in the ranking measure can be partly explained by the decreased number of condition pairs for each item.

3.4. Test 3

All of the 14 listeners in the test had participated in one of the two earlier tests, and thus they were all familiar with the task and user interface. The median age of the participants was 31 a, with the minimum 24 a and maximum 37 a.

Discussions with the participants after the test revealed that many of them had perceived the task more difficult than at the first time (meaning either Test 1 or Test 2). The listener comments indicated that this was caused by the large number of original items compared to the number of presented reverberation conditions for each item. This hindered the participants from creating a mental model of the original signals for internal reference. This effect is rather interesting as it suggests that the perceived level of reverberation is not dependent only on signal, but also expectations related to it. Despite this perceived difficulty, the listener performance analysis illustrated in Fig. 11 shows that the degradation in performance from earlier tests is almost negligible. The individual item ratings are not illustrated because of the condition subsampling.

Fig. 12 illustrates the mean standard deviation of listener ratings for each item, averaged over the conditions. While the majority of the values lie at approximately 12, "metal" item sticks out having the standard deviation of 20 points. This suggests that the listener agreement depends also considerably on the underlying signal.

4. DISCUSSION

Even though it was partly expected, the considerable rating differences caused by the source material raise a question. Specifically, in the first test, speech was consistently perceived as highly reverberant, while the orchestral music was perceived as low-reverberant for the same reverberation parameters. Similarly, the acoustic guitar item was perceived to be very low-reverberant.[1]. Adding the same amount of reverberation to signals of different

[1] Remember that all these three original signals were recorded in anechoic conditions.

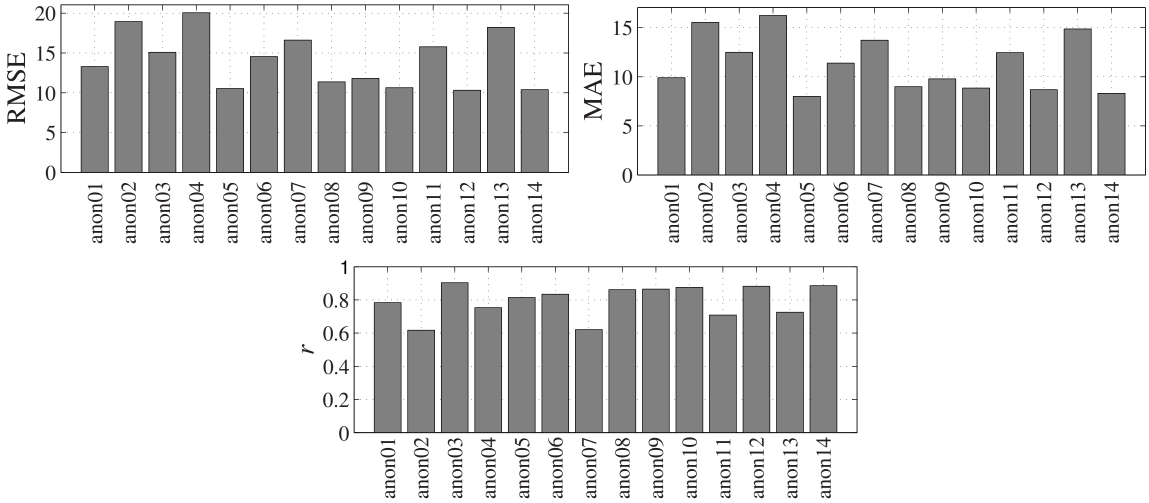

Figure 11: Test 3 Inter-listener evaluation results. For each participation, the evaluation measure is calculated between the participation and the mean of all other participations. Top left panel: RMSE, the mean value is 14.1. Top right panel: MAE, the mean value is 11.4. Bottom panel: correlation coefficient, the mean value is 0.80.

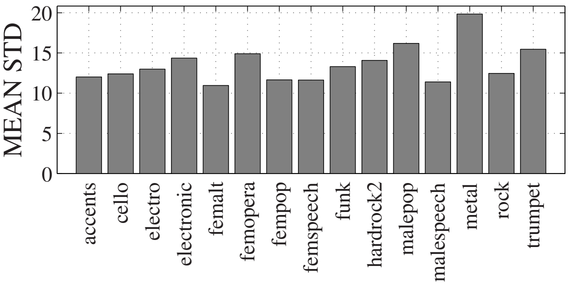

Figure 12: Test 3 Mean rating standard deviation for each test item.

classes thus causes different perceived reverberation levels.

It can be hypothesized that these signal class-dependent differences may be caused by two factors: either some physical property that is intrinsic to the signals and / or learned expectations which are associated by humans when listening to signal classes that they may have been listened to before in certain contexts (extrinsic factors). Discussions with the test participants suggested that the speech signal was perceived to be non-reverberant only when it was practically clear of all reverberation, and even the smallest amount of added reverberation was already perceived as prominent. This may be due to the fact that the shortest reverberation time in the experiment was 1.0 s, which is already longer than the reverberation time in normal room environments, and thus the signals were “more reverberant than usually expected” (extrinsic factor), as commented by several of the listeners. This expectation assumption is supported also by the results for the symphony orchestra: normally this type of stimulus is played under rather reverberant conditions (concert halls) with large portions of the actual signal heard being reverberation as an important aesthetic aspect.

On the other hand, the hypothesis of dependence on the physical properties of the signal (intrinsic factors) is highly plausible and supported by experimental evidence. Often the audibility of reverberation within a complex music signal is attributed to the amount of short term non-stationarity of the music signal leaving enough “gaps” between onsets or transient signal regions that reverberation contributions can become audible the the listener rather than being masked (see, e.g., [5, 6, 17]). While orchestral music contains numerous note onsets, the majority of the participating instruments are of non-percussive nature and the sharpness of note onsets and offsets are smeared by the variations in musicians’ timings that are inherent to a large ensemble. Both factors promote the orchestral music to be perceived as little reverberant. On the other hand, speech is known to be a signal with very high non-stationary. In fact, a char-

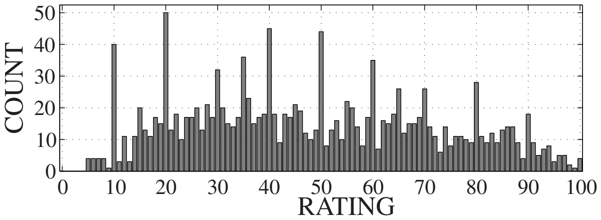

Figure 13: Test 1 Histogram of all the ratings illustrating the quantization to GUI tick line artefact.

acteristic modulation frequency of 4–5 Hz is considered typical for speech [1, 14]. With each change between vowel and non-vowel phonemes, the spectral balance between high and low frequency components in the signal changes significantly, leading to many time segments that can expose reverberant components.

In this work we have assumed that asking listening subjects for a rating of the perceived "level of reverberation" is a meaningful question in the sense that it carries a sufficiently well-defined meaning within the population of test listeners. Considering the obtained confidence intervals of the listener ratings, this assumptions appears to hold to a considerable extent. However, informal interviews with the participants revealed that many of them were analyzing different cues from the signal to determine the overall level. Some indicated to have rated the signals based on fast first impression, "a gut feeling", while some attempted to listen the signals more analytically. The analytical listeners could then be further divided into two groups: the ones trying to imagine the space based on the acoustic information and ground the rating to that, and the most analytical listeners dividing signal into the direct and reverberation components and basing their judgements on them.

As was noted in [18], the presence of the tick marks on the scale in the interface used for the rating causes quantization to those values. This artefact was expected also in this test, but still the prominence of the quantization was surprising, as can be seen in Fig. 13. The figure contains a histogram of all ratings from Test 1. The GUI contained tick marks on every 10 points, and it is possible to observe quantization to those values in the histogram. This should be taken into account when designing further tests.

5. CONCLUSIONS AND FUTURE WORK

This paper has presented extensive listening tests on the perceived level of reverberation in various audio signals. The results show that for an equal average loudness level, the shape of the reverberation tail is important for the level perceived: a longer reverberation tail is perceived to have a higher level even though it has a lower absolute instantaneous level. The source material has a considerable effect on the level perceived, leading to both an offset individual to each sound item, and an individual range of the ratings. The differences in ratings between individuals are close to the average difference between repeats of an individual and were similar in all tests. Somewhat unexpectedly, mono and stereo reverberations with equal reverberation time and level were found to be judged to have equal perceptual levels.

The presented investigation suggests several topics for potential future work. First of all, more thorough sampling of the reverberation physical parameter space would be of interest for establishing a more rigorous computational model mapping the mixing parameters to subjective space. For enhanced naturalness, it would be beneficial to include some model of early reflections in the impulse response. Finally, a large and interesting area for future investigations would be to study possible intrinsic and extrinsic factors affecting the perceived level of reverberation and their individual significance for explaining the observed measurements.

ACKNOWLEDGEMENTS

The authors would like to thank Sunish George for the fruitful discussions and feedback regarding the test design.

6. REFERENCES

[1] J. R. Ashley. Auditory backward integration can ruin a concert hall. In *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 330–334, Washington, D.C., USA, 1979.

[2] J. Blauert. *Spatial hearing: the psychophysics of human sound localization*. MIT Press, Cambridge, Mass., USA, revised edition, 1997.

[3] W. G. Gardner and D. Griesinger. Reverberation level matching experiments. In *Proc. of W. C. Sabine Centennial Symposium*, pages 263–266, Cambridge, Mass., USA, June 1994.

[4] S. George, S. Zielinski, F. Rumsey, P. Jackson, R. Conetta, M. Dewhirst, D. Meares, and S. Bech. Development and validation of an unintrusive model for predicting the sensation of envelopment arising from surround sound recordings. *Journal of the Audio Engineering Society*, 58(12):1013–1031, Dec. 2010.

[5] D. Griesinger. Further investigation into the loudness of running reverberation. In *Proc. of Institute of Acoustics Conference*, London, UK, Feb. 1995.

[6] D. Griesinger. How loud is my reverberation? In *Proc. of 98th Audio Engineering Society Convention*, Paris, France, Feb. 1995.

[7] International Telecommunication Union. *ITU-R BS.1534-1: Method for the Subjective Assessment of Intermediate Quality Level of Coding Systems*. Geneva, Switzerland, 2003.

[8] International Telecommunication Union. *ITU-R BS.1770-1: Algorithms to measure audio programme loudness and true-peak audio level*. Geneva, Switzerland, Dec. 2007.

[9] M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1–2):81–93, June 1938.

[10] J. C. R. Licklider and E. Dzendolet. Oscillographic scatterplots illustrating various degrees of correlation. *Science*, 107(2770):121–124, Jan. 1948.

[11] J. Merimaa, T. Peltonen, and T. Lokki. Concert hall impulse responses - Pori, Finland: Analysis results. Technical report, Helsinki University of Technology, May 2005.

[12] J. A. Moorer. About this reverberation business. *Computer Music Journal*, 3(2):13–28, June 1979.

[13] J.-D. Polack and H. Alrutz. Ein statistisches Modell der Impulsantwort eines Raumes. In *Fortschritte der Akustik - DAGA'84*, pages 403–406, Darmstadt, Germany, Mar. 1984. In German.

[14] E. Scheirer and M. Slaney. Construction and evaluation of a robust multifeature speech/music discriminator. In *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 1331–1334, Munich, Germany, 1997.

[15] S. Schlecht. Options and limits of feedback delay networks for artificial reverberation of audio signals - A time and frequency domain approach. Master's thesis, University of Trier, Trier, Germany, Oct. 2010.

[16] M. R. Schroeder. Natural sounding artificial reverberation. *Journal of the Audio Engineering Society*, 10(3):219–223, 1962.

[17] C. Uhle, J. Paulus, and J. Herre. Predicting the perceived level of late reverberation using computational models of loudness. In *Proc. of 17th International Conference on Digital Signal Processing*, Corfu, Greece, July 2011.

[18] S. Zieliński and F. Rumsey. On some biases encountered in modern audio quality listening tests – a review. *Journal of the Audio Engineering Society*, 56(8):427–451, Jun 2008.