OPEN ACCESS Freely available online

Journal of the Audio Engineering Society
Vol. 65, No. 1/2, January/February 2017
DOI: https://doi.org/10.17743/jaes.2016.0062

Perceptual Evaluation and Analysis of Reverberation in Multitrack Music Production

BRECHT DE MAN,[1] *AES Member*, **KIRK McNALLY,**[2] *AES Member*, **AND**
(b.deman@qmul.ac.uk) (kmcnally@uvic.ca)

JOSHUA D. REISS,[1] *AES Member*
(joshua.reiss@qmul.ac.uk)

[1]*Centre for Digital Music, Queen Mary University of London, London, UK*
[2]*University of Victoria, Victoria, BC, Canada*

Artificial reverberation is an important music production tool with a strong but poorly understood perceptual impact. A literature review of the relevant works concerned with the perception of musical reverberation is provided, and the use of artificial reverberation in multi-source mixes is studied. The perceived amount of total artificial reverberation in a mixture is predicted using relative reverb loudness and early decay time, as extracted from the newly proposed Equivalent Impulse Response. Results indicate that both features have a significant impact on the perception of a mix and that they are closely related to the upper and lower bounds of desired amount of reverberation in a mixture.

0 INTRODUCTION

Reverberation is one of the most important tools at the disposal of the audio engineer. Essential in any recording studio or live sound system [1], the use of artificial reverb (simply referred to as "reverb" in this work) is widespread in most musical genres and it is among the most universal types of audio processing in music production.

Despite its prominence in music production, there are few studies on the usage and perception of artificial reverberation relevant to this context. The limited research may relate to a lack of universal parameters and interfaces, while algorithms across the available reverb units vary wildly. In comparison, typical equalization (EQ) parameters are standardized and readily translate to other implementations.

The ability to predict the desired amount of reverberation with a reasonable degree of accuracy has applications in automatic mixing and intelligent audio effects [2, 3], novel music production interfaces (e.g., various mappings of low-level parameters to more perceptually relevant parameters or terms [4, 5]), and compensation of listening conditions [6].

In this work, the previous studies concerned with the automation, preference, and perception of reverberation in music are critically reviewed to establish the requirements for a new methodology (Sec. 1). The problem and definitions used in the remainder of the work are established in Sec. 2. Sec. 3 presents an experiment where a dataset of mixes is perceptually evaluated to explore the relationship between perceived amount of reverberation and the underlying objective parameters. Analysis of the annotated subjective responses is discussed in Sec. 4. In Sec. 5, the ITU-R BS.1770 loudness of the reverb versus that of the "direct sound" is tested against the mix evaluations. Then, the concept of an *Equivalent Impulse Response* is introduced and its reverberation time is assessed as a predictor of perceived amount of reverberation (Sec. 6). Concluding remarks and a discussion of future work ensue in Secs. 7 and 8.

1 BACKGROUND

In contrast to other important mix engineering tools—such as level [7, 8], panning [9], EQ [10], and dynamic range compression [11, 12]—to date only one attempt at automatic control of reverberators has been made [3]. Very little work is available on novel, more intuitive interfaces for reverb [13, 14] and mapping terms to its parameters [4, 5]. A number of studies have looked at perception of reverberation in musical contexts [2, 6, 15–32], see Table 1.

The focus of this study is the perception of artificial reverberation of multi-source materials taken from examples of fully-realized, professional music productions. The present case stands apart from the work cited above, where the effect of reverb parameters on the subject's preference or perception is under investigation, as applied to a single source, and typically isolated from any musical, visual or sonic context. As reverberation is a complex and multifaceted matter, controlled experiments are often required. Several of these studies involved only a single, simple, and potentially unpleasant and unfamiliar reverberator

Table 1. Overview of studies concerning perception of reverberation of musical signals. Test method: PE or DA (Perceptual Evaluation or Direct Adjustment of reverb settings); participants Skilled or Unskilled in audio engineering. Reverberator properties: Stereo or Mono; Early Reflections or No Early Reflections.

		Stereo		Mono	
		ER	No ER	ER	No ER
PE	Skilled	[15, 24, 31, 32]		[18, 30]	[20]
	Unskilled	[16, 19, 21–23]	[2]	[6]	[17, 25]
DA	Skilled	[32]	[28]	[26]	[27]
	Unskilled	[22, 29]			[25]

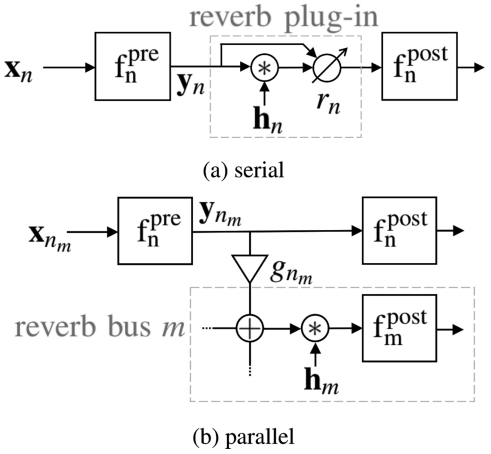


Fig. 1. Reverb signal chains.

[15, 16], sometimes without the use of early reflections [2, 17] or stereo capabilities [6, 18]. In some cases, the number of reverberator parameters were limited, often taking a restricted range or set of values [19–21], and applied to a single (type of) source sample [22–24]. In [3, 25] the parameter values considered were set by unskilled participants using unfamiliar tools and inferior listening environments. Finally, the results of several parameter adjustment tests are not validated through perceptual evaluation [26–28].

It has not yet been investigated whether the perception of reverberation amount and time of a single source in isolation has any relevance within the context of multitrack music production, inherently a multidimensional problem, where different amounts and types of reverb are usually applied to different sources, which are then combined to form a coherent mixture. Thus, while relevant for the respective studies, these works may not offer insight into how an audio professional might use reverb in a commercial music production environment.

In order to better understand the use, perception, and preference with regards to reverberation in music, it is deemed necessary to study its application by trained engineers using familiar, professional grade tools in the context of a complete, representative mix. The results of such application should be subjectively evaluated to validate the engineers' choices and gain additional insight into the perceptual impact of differences in reverb. The methodology presented herein, along with the findings from a particular dataset, accommodates analysis of practice and perception of reverb in a less controlled, ecologically valid setting.

2 PROBLEM FORMULATION

In what follows, the perceived amount of reverberation is predicted based on objective features extracted from both the combined reverb signal and the remainder of the mix. These signals will be referred to as wet ($\mathbf{s}_{wet}$) and dry ($\mathbf{s}_{dry}$), respectively. They are not always easy to extract in practice, even when all source audio and DAW session files, including all parameter settings, are available. This is due to the following conditions:

1) Different amounts and types of reverb are applied to the different sources in the mixture; and
2) Post-reverb, nonlinear processing (dynamic range compression, fader riding, automation of parameters) as well as linear processing (weighting, EQ) are applied to the individual sources as well as the complete mix or subgroups thereof.

Omitting time arguments for readability, tracks $n = 1, \ldots, N$ carry the source signals $\mathbf{x}_n$ that are often already processed before any reverb is applied, giving $\mathbf{y}_n = \mathrm{f}_\mathrm{n}^\mathrm{pre}(\mathbf{x}_n)$. Reverb (with impulse response $\mathbf{h}_n$) can be added to the processed tracks $\mathbf{y}_n$ using serial processing, with the reverb plug-in inserted "in-line," where the gain ratio $r_n \in [0, 1]$ between the wet and dry signal is set within the plug-in (Fig. 1a). Alternatively, reverb is added through parallel processing, with tracks scaled by a gain factor g and sent to a reverb plug-in on a separate bus. Typically, several tracks $n_m = 1, \ldots, N_m$ are sent to the same reverb bus m (Fig. 1b). In both cases, further processing $\mathrm{f}_\mathrm{n}^\mathrm{post}(\cdot)$ is then applied to the respective tracks and buses, i.e., post-reverb. The wet and dry part of the mix can therefore be expressed as:

$$\mathbf{s}_{wet} = \sum_n \mathrm{f}_\mathrm{n}^\mathrm{post}(r_n \mathbf{h}_n * \mathbf{y}_n) + \sum_m \mathrm{f}_\mathrm{m}^\mathrm{post}\left(\mathbf{h}_m * \sum_{n_m} g_{n_m}\mathbf{y}_{n_m}\right) \quad (1)$$

$$\mathbf{s}_{dry} = \sum_n \mathrm{f}_\mathrm{n}^\mathrm{post}((1 - r_n)\,\mathbf{y}_n) \quad (2)$$

With $\mathbf{h}'_n = (r_n \mathbf{h}_n + (1 - r_n)\,\delta)$ as the total impulse response of the in-line reverb, reverberant ratio r_n included, where δ is the unit impulse, the total mix $\mathbf{s}_{tot}$ then becomes:

$$\mathbf{s}_{tot} = \sum_n \mathrm{f}_\mathrm{n}^\mathrm{post}(\mathbf{h}'_n * \mathbf{y}_n) + \sum_m \mathrm{f}_\mathrm{m}^\mathrm{post}\left(\mathbf{h}_m * \sum_{n_m} g_{n_m}\mathbf{y}_{n_m}\right) \quad (3)$$

which is equal to $\mathbf{s}_{dry} + \mathbf{s}_{wet}$ as long as the condition $\mathrm{f}_\mathrm{n}^\mathrm{post}(a + b) = \mathrm{f}_\mathrm{n}^\mathrm{post}(a) + \mathrm{f}_\mathrm{n}^\mathrm{post}(b)$ is satisfied. For this to be true, post-reverb nonlinear processing $\mathrm{f}_\mathrm{n}^\mathrm{post}(\cdot)$ is applied to both the wet and dry signal in such a way that their sum still equals the original mix. Any gain changes applied by a dynamic range compressor are dependent on

its side-chain signal (equal to the input signal by default). The original mixed signal is thus used for this side-chain signal when processing the dry or wet signal. In other words, in Eqs. (1) and (2), $\mathrm{f}_{\mathrm{n}}^{\mathrm{post}}(\cdot) = \mathrm{f}_{\mathrm{n}}^{\mathrm{post}}\left(\cdot, \mathbf{h}'_n * \mathbf{y}_n\right)$, the extra argument representing the side-chain signal, so that $\mathbf{s}_{tot} \equiv \mathbf{s}_{dry} + \mathbf{s}_{wet}$. For simplicity, it is assumed that this post-processing is applied per track, though in reality it can be applied to groups of sources simultaneously.

The interest herein is how the perceived excess or lack of reverberation amount is influenced by the difference between the loudness of the reverb signal and the dry signal (see [2, 6, 32]), as well as the overall reverberation time (see [2, 15, 24]).

The first considered feature, relative reverb loudness (RRL), is defined as:

$$\mathrm{RRL} = \overline{\mathrm{ML}\left(\mathbf{s}_{wet}\right) - \mathrm{ML}\left(\mathbf{s}_{dry}\right)} \tag{4}$$

where ML is the Momentary Loudness in loudness units (LU) as specified in [33]. The difference of the momentary loudness of the wet and dry signal is calculated for each measurement window, and the average ($\overline{x}$) is taken over each window. It should be noted that (forward) masking and binaural dereverberation are not taken into account with this measure. More advanced partial loudness features were used in [2] to predict the perceived amount of reverb. However, such features[1] were not used in this work because the authors found they did not perform well on the considered content, showing weak correlation with perception, and more work is needed to establish the applicability of multi-band loudness models [34], specifically to multi-source music. Furthermore, the simple filtered RMS measure used here is far less computationally expensive and suitable for real-time applications.

The second feature, reverberation time, is usually derived from the reverberation impulse response (RIR). In the context of this study, however, the RIR is not readily defined, due to conditions (1) and (2) above. As such, the transformation between the mix *without* reverb and the mix *with* reverb is not a linear one, and it cannot be defined by an impulse response, even if the reverberator used is applying a linear transformation (which is also not always the case [35]). However, an Equivalent Impulse Response (EIR) can be estimated in which temporal and spectral aspects of the total reverb are embedded:

$$\mathbf{s}_{wet} \approx \mathbf{h}_{eq} * \mathbf{s}_{dry} \tag{5}$$

From such an impulse response, traditional (acoustic) reverberation parameters can be extracted, which describe the overall reverberation in universally defined terms such as reverberation time, along with clarity, IR spectral centroid, and central time, which can then be translated to other reverberators [4].

3 METHOD

3.1 Design

A set of mixes was created for a number of songs and subsequently compared against each other and subjectively assessed in a multiple-stimulus test. The mixes were to be rated according to "preference" as well as commented on with a free-form text response. The preference rating serves to determine the overall appreciation of the mix and how this correlates with audio features extracted from the mix and its components (see [36]). It further forces the subject to consider which mix they prefer over which, so that they reflect and comment on the aspects that have an impact on their preference.

The goal of this experiment was to uncover which mixes were spontaneously perceived as too reverberant or as not reverberant enough. Therefore, the subjects were not explicitly asked to rate the perceived amount of reverberation. Rather, analysis of the free-form comments reveals mixes in which reverberation—and the relative lack or abundance thereof—was referenced as an issue.

The independent variables of the experiment were mix (or mix engineer) and song. The dependent variables consisted of the preference rating and the free-choice profiling results.

3.2 Participants

The mixes were created by 24 master level sound recording students from the same program, all musicians with a Bachelor of Music degree. Each song was mixed by a group of eight students, where each individual student mixed between one and five songs. The average participant was 25.1 $\pm$ 1.8 years old, with 5.1 $\pm$ 1.9 years of audio engineering experience. Of the 24 participants, 5 were female and 19 were male.

For the perceptual evaluation experiment there were a total of 34 participants: 24 participants from the mix creation process and 10 instructors from the same sound recording program. For each individual song, between 12 and 16 subjects assessed the different mixes. In the context of this work, students did not evaluate any songs they had previously mixed. Each student received a small compensation for their time upon taking part in the listening test.

3.3 Materials

Multitrack recordings of 10 different songs, played by professional musicians and recorded by Grammy award-winning recording engineers, were given to the students tasked with creating a stereo mix from the source tracks. A total of 80 student mixes were created for the experiment. With a few exceptions, the students were unfamiliar with the content before the experiment. Table 2 lists all songs used in the experiment. Those which have a Creative Commons (CC) license have been made available on the Open Multitrack Testbed[2] [37], including source tracks and mixes.

[1] github.com/deeuu/loudness/

[2] multitrack.eecs.qmul.ac.uk

Table 2. Songs used in the experiment.

	Song	Artist	CC
1	In The Meantime	Fredy V	√
2	Lead Me	The DoneFors	√
3	My Funny Valentine	Joshua Bell, Kristin Chenoweth	
4	No Prize	Dawn Langstroth	
5	Not Alone	Fredy V	√
6	Pouring Room	The DoneFors	√
7	Red To Blue	Broken Crank	√
8	Under A Covered Sky	The DoneFors	√
9	Artist A[3]	Song A[3]	
10	Artist B[3]	Song B[3]	

A constrained but representative set of software tools was used to create the mixes, consisting of an industry standard digital audio workstation (DAW) with standard native plug-ins and additional professional reverb plug-ins. The students were familiar with all of these tools. Restricting the toolset allowed for extensive analysis of parameters and the ability to recreate the mix or its constituent tracks, with the various processing units enabled or disabled. As such, the reverb signals could be isolated from the rest of the mix.

The participants produced the different mixes in their preferred mixing location, so as to achieve a natural and representative spread of environments without a bias imposed by a specific acoustic space, reproduction system, or playback level. A limit of six hours of mixing time was imposed on the participants, but no further directions were given.

In addition to these eight mixes, the original, commercial mix was also provided in the listening test, and in some cases a machine-made mix though these are not included in the analysis as the parameter data is not available for these versions. The songs were selected from a wide range of genres to average out differences in genre-specific mixing approaches and signal characteristics and to allow for analysis of the influence of genre.

Further analysis of the mixes (Secs. 5 and 6) was conducted using the 71 mixes where all parameters were accessible and the mix could be perfectly recreated. In the other cases, participants used more than the permitted set of tools.

3.4 Apparatus

The listening test interface (from [38, 39], see Fig. 2) consisted of a single horizontal preference axis, with each mix represented by a numbered, vertical marker, and a corresponding text box for comments on that mix. An extra text box was provided for general comments on all mixes or the song as a whole. No anchors or references were included, and each fragment could be auditioned as many times as desired. Song and mix order was fully randomized, and all mixes were scaled to equal loudness according to [40]. At the end of the fragment, playback would loop to the start of that fragment. The fragments were aligned so that upon switching between fragments, the new fragment would start playing from the corresponding position. Playback could be paused and reset to the beginning by clicking the stop button.

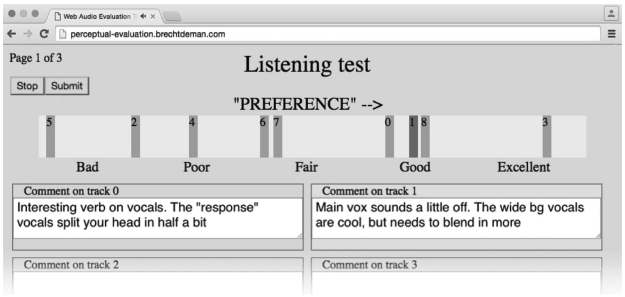


Fig. 2. Listening test interface.

The test took place in a professional-grade listening room with a high quality audio interface and loudspeakers [36]. Headphones were not used to avoid the sensory discrepancy between vision and hearing, as well as the expected differences in terms of preferred reverberation between headphone and speaker listening [41].

3.5 Procedure

The listening test was conducted with one participant at a time. After having been shown how to operate the interface, the participants were asked—both written and verbally—to audition the samples as often as desired, rate the different mixes according to their preference, and to write extensive comments in support of their ratings, for instance "why they rated a fragment the way they did" and "what was particular or different about it." They were instructed to first set the listening level as they wished, since their judgments are most relevant when listening at a comfortable and familiar level [42], and since the perceived reverberation amount varies with level [6, 25]. The instructions further stated participants could use the preference rating scale however they saw fit.

To reduce strain on the subjects, a fragment containing the second verse and second chorus of the song was selected from each mix, averaging one minute in length. This section was considered maximally representative as most sources were active in this part of the song. With up to 10 mixes per song, and up to 4 songs per test, the test length was well below the recommended duration limit of 90 minutes [43], and the possibility to take breaks was given to participants.

4 COMMENT ANALYSIS

To allow quantitative processing, every comment was split into its constituent statements. In total, 4227 separate statements were annotated from 1326 comments. Of these comments, 35.44% mention reverberation, and reverberation is not commented on by anyone in only 2 of the 80 mixes considered here. Furthermore, every subject commented on reverberation for at least 10% of the mixes they assessed. The comments were classified

[3]For two songs permission to disclose artist and song name was not granted.

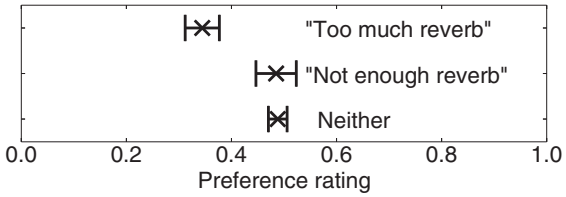


Fig. 3. Preference (0.0–1.0) per class: 95% confidence intervals.

into three classes: *"Too much reverb," "Not enough reverb,"* and—when unrelated to the perceived amount of reverberation—*Neither*.

Participants disagreed on whether there was too much or too little reverberation in only 4 of the 525 comments that mention reverberation. This supports the idea that mix engineers have a consistent judgment on the "correct" reverberation amount for a given mix. The low variance in the results may be explained by the fact that test participants are skilled listeners [25]. In the following sections, only comments regarding the subjective excess or shortage of reverberation of the whole mix (i.e., not any particular instrument) are considered.

Fig. 3 shows the mean preference ratings associated with statements from the different classes. As previously observed in [32, 44], the preference rating for a mix the subject found too reverberant is significantly lower than if it was considered too dry.

5 RELATIVE REVERB LOUDNESS

The relative reverb loudness is shown for each mix in Fig. 4, along with the number of subjects who indicated the mix was perceived as too reverberant or not reverberant enough, divided by the total number of subjects for that song. As expected, the majority of the mixes labelled "too reverberant" have a significantly higher relative reverb loudness than those labelled "not reverberant enough."

Overall, the preferred reverb loudness seems to differ significantly from [32], where the optimal reverb return loudness is estimated to be at −9 LU. In the current experiment, every mix with a relative reverb loudness of −9 LU or higher was judged to be too reverberant, and −14 LU appears to be a more desirable loudness as it is in between 95% confidence intervals of the medians of either labelled group.

The differences in reverb loudness are mostly subtle, with the just-noticeable difference (JND) of direct-to-reverberant ratio estimated at 5–6 dB [45], proof of the critical nature of the engineer's task. Despite this, there is a large level of agreement with regard to what mixes have a reverb surplus or deficit. The variance of preferred reverb level is considerably larger in [25], possibly due to the unskilled listeners.

There are some cases where despite a relatively high reverb loudness, subjects agreed that there was not enough reverberation (e.g., mix 3C or 5C in Fig. 4), or where mixes with a perceived excess of reverb did not exhibit a significantly higher-than-average measured loudness (e.g., 1B, 8P). Closer study of these outliers, through informal listening and analysis of parameter settings, revealed that mixes with a high perceived amount of reverberation but low measured reverb loudness typically have a long reverberation tail. Those marked as too dry have a strong, yet short and clear reverb signal, to the point of sounding similar to the dry input. As in [2], it would seem relative loudness of the reverb signal alone is generally insufficient to predict the perceived or preferred amount of reverberation. It is therefore believed that measuring the reverberation time will help explain the perceived amount of reverberation [21, 23, 31].

6 EQUIVALENT IMPULSE RESPONSE

6.1 Process

For the practical measurement of the EIR $\mathbf{h}_{eq}$ (see Eq. (5)) it is not possible to use sine sweep or maximum length sequence (MLS) methods due to condition (1) from Sec. 2. In the frequency domain, if $\mathrm{f}_\mathrm{n}^{\mathrm{post}}(\cdot)$ is a linear filter with frequency response $\mathbf{F}_n^{(post)}$, spectral division of the Fourier transforms of Eqs. (1) and (2) yields an equivalent frequency response:

$$\begin{aligned}\mathbf{H}_{eq} &= \frac{\mathbf{S}_{wet}}{\mathbf{S}_{dry}} \\ &= \frac{\sum_n \mathbf{F}_n^{(post)} r_n \mathbf{H}_n \mathbf{Y}_n + \sum_m \mathbf{F}_m^{(post)} \mathbf{H}_m \left(\sum_{n_m} g_{n_m} \mathbf{Y}_{n_m}\right)}{\sum_n \mathbf{F}_n^{(post)} (1 - r_n) \mathbf{Y}_n}\end{aligned} \tag{6}$$

In this case, the equivalent frequency response $\mathbf{H}_{eq}$ is a frequency- and gain-weighted version of the various reverb frequency responses $\mathbf{H}_n$ and $\mathbf{H}_m$, being dependent on the post-processing, the (pre-processed) input signals, and the wet to dry ratios. This interpretation is violated to the extent that $\mathrm{f}_\mathrm{n}^{\mathrm{post}}(\cdot)$ is not a linear function, see condition (2) from Sec. 2. In the case it is approximately linear but not stationary, the equivalent frequency response can describe the total reverb with reasonable accuracy as a function of time.

Neglecting any nonlinearities, the EIR is obtained by division of the signals ($\mathbf{s}_{wet}$ and $\mathbf{s}_{dry}$) in the spectral domain (also dual channel FFT analysis) [46]. Following Welch's method, complex averaging is performed on both the dry signal's power spectrum or auto spectrum ($\mathbf{G}_{dry,dry}^{(i)}$) and the cross spectrum ($\mathbf{G}_{dry,wet}^{(i)}$), taken from signal segments $i = 1 \ldots I$, with 50% overlap and a Hann window:

$$\begin{aligned}\mathbf{G}_{dry,dry}^{(i)} &= \mathbf{S}_{dry}^{*(i)} \mathbf{S}_{dry}^{(i)} \\ \mathbf{G}_{dry,wet}^{(i)} &= \mathbf{S}_{dry}^{*(i)} \mathbf{S}_{wet}^{(i)} \\ \mathbf{H}_{eq} &= \frac{\frac{1}{I}\sum_{i=1}^{I} \mathbf{G}_{dry,wet}^{(i)}}{\frac{1}{I}\sum_{i=1}^{I} \mathbf{G}_{dry,dry}^{(i)}} \equiv \frac{\mathbf{G}_{dry,wet}}{\mathbf{G}_{dry,dry}} \\ \mathbf{h}_{eq} &= \mathrm{iFFT}\left(\mathbf{H}_{eq}\right) = \mathrm{iFFT}\left(\frac{\mathbf{G}_{dry,wet}}{\mathbf{G}_{dry,dry}}\right)\end{aligned} \tag{7}$$

where iFFT is the inverse Fast Fourier Transform.

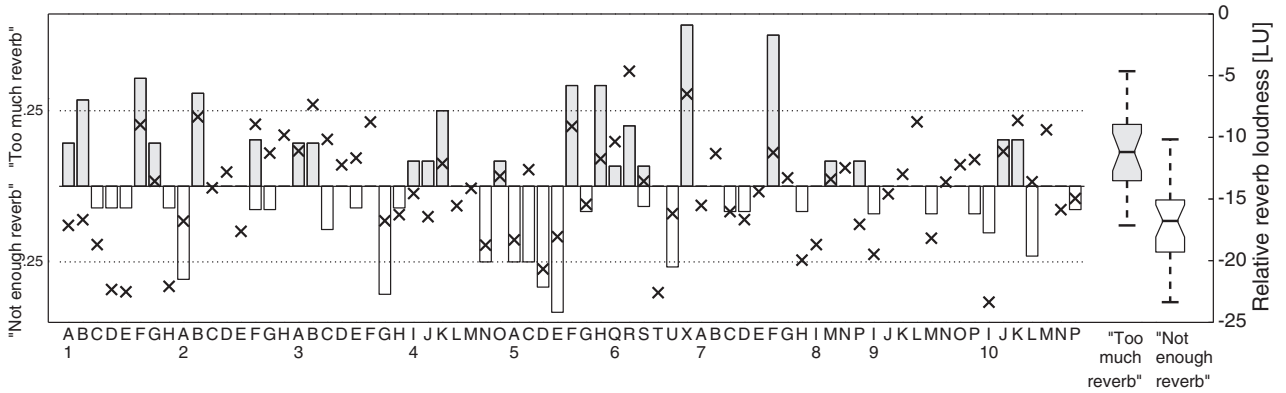


Fig. 4. Proportion of subjects who noted an excess or deficit of reverberation (bars), versus the relative loudness of the reverb signal (Xes). Letters denote different mix engineers, numbers denote different songs (see Table 2). The box plots show the relative loudness values for mixes collectively found to be too "wet" and "dry," respectively; here, the center line denotes the median, the box extends from the 25th to the 75th percentile, the notch is the median's confidence interval, and the whiskers span from the lowest to the highest value.

The window length has been empirically obtained to produce the impulse response with the lowest noise floor while still being sufficiently long compared to the reverberation times.

In contrast to most work on impulse response estimation and room impulse response inversion, in this case there is no reference or error measure to objectively evaluate the quality of the obtained impulse response. Convolving the dry signal with the EIR will rarely approximate the wet signal, due to condition (1).

While stereo reverberation generated from a monaural source is generally defined by two impulse responses (one for each channel), and stereo reverberation of a stereo source by four ($\mathbf{h}_{L\to L}$, $\mathbf{h}_{L\to R}$, ...), for the purpose of this study a single impulse response is extracted from the spectral division of the wet and dry signal, each summed to mono. It has been shown that with identical reverberation times and level, mono and stereo reverberation signals are perceived as having equal loudness regardless of the source material [44].

From this impulse response, it is possible to extract reverberation time measures such as the Early Decay Time (EDT). This is a particularly suitable feature as the calculated impulse responses are noisy. Furthermore, it has been shown that the EDT is more closely related to the conscious perception of reverberation, especially while the source is still playing during the reverberation decay, as is the case here [14, 31].

6.2 Equivalent Impulse Response Analysis and Results

Fig. 5 shows all mixes as a function of their reverb loudness and reverb time and labeled according to the net number of subjects who classified them as either *"Too much reverb," "Not enough reverb,"* or *Neither*. The relative reverb loudness is as computed in Sec. 5, and the EDT is calculated from the EIR using the decay method, equivalent to six times the time it takes for the decay curve to reach −10 dB, an estimation of T_{60} [47]. The logarithm of

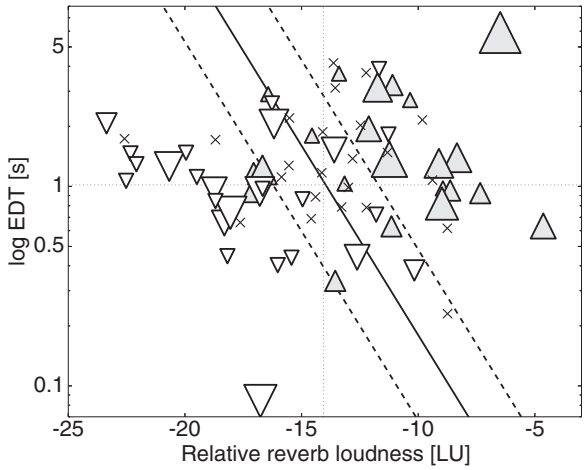


Fig. 5. Mixes where subjects noted an excess (grey upwards triangle) or deficit (white downwards triangle) of reverb, or neither (X), as a function of the relative reverb loudness and the EDT of the reverb signal. Marker size is scaled by net number of subjects, and logistic regression decision boundaries are shown.

the EDT is used to better visualize a few large values, and this also makes the distribution normal.

As the dependent variable is a binary classification into "too reverberant" or "not reverberant enough," a logistic regression is performed based on the measurements of relative reverb loudness and EDT, for each assignment to either category by a subject. Comparing this to a restricted model with only the relative reverb loudness (RRL) as a predictor variable, a statistically significant increase is seen in the model fit (likelihood ratio $-2\ln L_{both}/L_{RRL} = 7.749$, i.e. $p = .005$ on a χ^2 distribution)—that is, the EDT is indeed helpful in explaining the perception of the reverberation amount. The decision boundaries at .25, .50, and .75 are shown in Fig. 5, along with the .50 decision boundaries for the individual predictor variables.

Such a sharp transition between what is considered too reverberant and too dry, again emphasizes the importance of careful adjustment of reverb parameters. This is further supported by the observation in [29] that masking causes

Table 3. Logistic regression results.

	Coeff	SE	$P > \|z\|$	95% CI
RRL	0.4866	0.089	0.000	0.312 – 0.662
EDT	2.5619	1.043	0.014	0.519 – 4.605
Intercept	6.7767	1.282	0.000	4.263 – 9.290

reverberation audibility to decrease by 4 dB for every dB decrease in reverberant level. The differences in reverberation time between the different mixes are mostly of the order of the JND [18], as was the case with the differences in relative reverb loudness.

7 SUMMARY AND CONCLUSION

An experiment was conducted where 80 mixes were generated from 10 professional-grade music recordings by trained engineers in a familiar and commercially representative setting, which were then rated in multi-stimulus listening tests. Annotated subjective comments were analyzed to determine the importance of reverberation in the perception of mixes, as well as to classify mixes having too much or too little overall reverberation. This study is different from previous work in that it examines reverb in a relevant music production context, where reverb is applied to multiple tracks in varying degrees and types.

Although the perceptual evaluation experiment purposely did not mention reverberation as a feature to consider, it is commented on in 35% of the cases, confirming that differences in reverb use have a large impact on the perceived quality of a mix [44], as assessed by skilled listeners. Notwithstanding the less controlled nature of the study, variance in its findings is significantly narrower than in similar work, likely due in part to proficiency of participants in both the mix experiment and subsequent perceptual evaluation.

To a large extent, the relative reverb loudness gives a suitable indication of how audible or objectionable reverberation is. These subjective judgments are further predicted by considering reverb decay time, derived from a newly proposed Equivalent Impulse Response that captures reverberation characteristics for a mixture of sources with varying degrees and types of reverb. Both measures are suitable for real-time applications such as automated reverberators or assistive interfaces.

The results support the notion that a universally preferred amount of reverberation is unlikely to exist, but show that upper and lower bounds can be identified with reasonable confidence. The importance of careful parameter adjustment is evident from the limited range of acceptable feature values with regard to perceived amount of reverberation, when compared to the just-noticeable differences in both relative reverb loudness and the Equivalent Impulse Response's EDT. This study confirms previous findings that a perceived excess of reverberation typically has a more detrimental effect on subjective preference than when the reverberation level was indicated to be too low, suggesting it is better to err on the "dry" side.

8 FUTURE WORK

Future implementations should take into account how reverberant the "dry" signal is, particularly when the original tracks contain a significant amount of reverberation. Source separation or dereverberation could help separate the two for a more accurate estimation of the dry and wet sound.

A new dataset with mixes and perceptual evaluations from subjects of various backgrounds, locations, and levels of expertise (including laymen) is required in order to analyze the consistency of reverberation preferences across different populations.

Artificial reverberation is defined by far more attributes, objective and perceptual, than those covered in this work. Further features and parameters to consider include predelay [29], echo density [35], autocorrelation [32], and more sophisticated loudness features [2].

Finally, the data collected in this mix experiment and the subsequent perceptual evaluation can be used to study perception and use of other music production tools such as balance, EQ, and dynamic range compression. In the interest of reproducibility and to allow easy extension of this work, the source tracks, stereo mixes, DAW files, and extracted reverberant and dry signals were made available in the Open Multitrack Testbed[4] [37] for the six songs licensed under a Creative Commons license.

9 ACKNOWLEDGMENTS

This work was made possible by the Engineering and Physical Sciences Research Council Grant EP/K009559/1 "Platform Grant: Digital Music," 2013–18. The authors also wish to thank Dominic Ward for a fruitful discussion on loudness models and related features.

10 REFERENCES

[1] B. A. Blesser, "An Interdisciplinary Synthesis of Reverberation Viewpoints," *J. Audio Eng. Soc.*, vol. 49, pp. 867–903 (2001 Oct.).

[2] C. Uhle et al., "Predicting the Perceived Level of Late Reverberation Using Computational Models of Loudness," *17th Int. Conf. on DSP*, pp. 1–7 (2011 July). https://doi.org/10.1109/ICDSP.2011.6004990

[3] E. T. Chourdakis and J. D. Reiss, "A Machine Learning Approach to Application of Intelligent Artificial Reverberation," *J. Audio Eng. Soc.*, vol. 65, pp. 56–65 (2017 Jan./Feb.). https://doi.org/10.17743/jaes.2016.0069

[4] Z. Rafii and B. Pardo, "Learning to Control a Reverberator Using Subjective Perceptual Descriptors," *10th ISMIR Conf.* (2009 Oct.).

[5] R. Stables et al., "SAFE: A System for the Extraction and Retrieval of Semantic Audio Descriptors," *15th ISMIR Conf.* (2014 Oct.).

[6] C. Bussey et al., "Metadata Features that Affect Artificial Reverberator Intensity," presented at the *AES 53rd*

[4] multitrack.eecs.qmul.ac.uk

International Conference: Semantic Audio (2014 Jan.), conference paper P2-10.

[7] D. Dugan, "Automatic Microphone Mixing," *J. Audio Eng. Soc.*, vol. 23, pp. 442–449 (1975 July/Aug.).

[8] E. Perez Gonzalez and J. D. Reiss, "Automatic Gain and Fader Control for Live Mixing," *IEEE WASPAA* (2009 Oct.). https://doi.org/10.1109/ASPAA.2009.5346498

[9] E. Perez Gonzalez and J. D. Reiss, "A Real-Time Semiautonomous Audio Panning System for Music Mixing," *EURASIP J. Adv. Sig. Pr.* (2010 May). https://doi.org/10.1155/2010/436895

[10] S. Hafezi and J. D. Reiss, "Autonomous Multitrack Equalization Based on Masking Reduction," *J. Audio Eng. Soc.*, vol. 63, pp. 312–323 (2015 May). https://doi.org/10.17743/jaes.2015.0021

[11] D. Giannoulis et al., "Parameter Automation in a Dynamic Range Compressor," *J. Audio Eng. Soc.*, vol. 61, pp. 716–726 (2013 Oct.).

[12] Z. Ma et al., "Intelligent Multitrack Dynamic Range Compression," *J. Audio Eng. Soc.*, vol. 63, pp. 412–426 (2015 June). https://doi.org/10.17743/jaes.2015.0053

[13] P. Seetharaman and B. Pardo, "Crowdsourcing a Reverberation Descriptor Map," *ACM Int. Conf. on Multimedia* (2014 Nov.). https://doi.org/10.1145/2647868.2654908

[14] J.-M. Jot and O. Warusfel, "Spat∼: A Spatial Processor for Musicians and Sound Engineers," *CIARM: Int. Conf. on Acoustics and Musical Research* (1995 May).

[15] A. Czyzewski, "A Method of Artificial Reverberation Quality Testing," *J. Audio Eng. Soc.*, vol. 38, pp. 129–141 (1990 Mar.).

[16] Y. Ando et al., "On the Preferred Reverberation Time in Auditoriums," *Acta Acustica united with Acustica*, vol. 50, pp. 134–141 (1982 Feb.).

[17] I. Frissen et al., "Effect of Sound Source Stimuli on the Perception of Reverberation in Large Volumes," *Auditory Display: 6th Int. Symposium*, pp. 358–376 (2010 May). https://doi.org/10.1007/978-3-642-12439-6_18

[18] Z. Meng et al., "The Just Noticeable Difference of Noise Length and Reverberation Perception," ISCIT, pp. 418–421 (2006 Oct.). https://doi.org/10.1109/ISCIT.2006.339980

[19] A. H. Marshall et al., "Acoustical Conditions Preferred for Ensemble," *J. Acoust. Soc. Amer.*, vol. 64, pp. 1437–1442 (1978 Nov.). https://doi.org/10.1121/1.382121

[20] P. Luizard et al., "Perceived Suitability of Reverberation in Large Coupled Volume Concert Halls," *Psychomusicology*, vol. 25, p. 317–325 (2015 Sep.). https://doi.org/10.1037/pmu0000109

[21] S. Hase et al., "Reverberance of an Existing Hall in Relation to Both Subsequent Reverberation Time and SPL," *J. Sound. Vib.*, vol. 232, pp. 149–155 (2000 Apr.). https://doi.org/10.1006/jsvi.1999.2690

[22] M. Barron, "The Subjective Effects of First Reflections in Concert Halls—The Need for Lateral Reflections," *J. Sound Vib.*, vol. 15, pp. 475–494 (1971 Apr.). https://doi.org/10.1016/0022-460X(71)90406-8

[23] G. A. Soulodre and J. S. Bradley, "Subjective Evaluation of New Room Acoustic Measures," *J. Acoust. Soc. Amer.*, vol. 98, pp. 294–301 (1995 July). https://doi.org/10.1121/1.413735

[24] M. R. Schroeder et al., "Comparative Study of European Concert Halls: Correlation of Subjective Preference with Geometric and Acoustic Parameters," *J. Acoust. Soc. Amer.*, vol. 56, pp. 1195–1201 (1974 Oct.). https://doi.org/10.1121/1.1903408

[25] J. Paulus et al., "A Study on the Preferred Level of Late Reverberation in Speech and Music," presented at the *AES 60th International Conference: DREAMS (Dereverberation and Reverberation of Audio, Music, and Speech)* (2016 Jan.), conference paper 1-3.

[26] D. Lee and D. Cabrera, "Equal Reverberance Matching of Music," *Proc. Acoustics* (2009 Nov.).

[27] D. Lee et al., "Equal Reverberance Matching of Running Musical Stimuli Having Various Reverberation Times and SPLs," *20th ICA* (2010 Aug.).

[28] W. G. Gardner and D. Griesinger, "Reverberation Level Matching Experiments," *Sabine Centennial Symposium* (1994 June).

[29] D. Griesinger, "How Loud Is My Reverberation?" presented at the *98th Convention of the Audio Engineering Society* (1995 Feb.), convention paper 3943.

[30] W. Kuhl, "Über Versuche zur Ermittlung der günstigsten Nachhallzeit großer Musikstudios," *Acta Acustica united with Acustica*, vol. 4, pp. 618–634 (1954 Jan.).

[31] E. Kahle and J.-P. Jullien, "Some New Considerations on the Subjective Impression of Reverberance and its Correlation with Objective Criteria," *Sabine Centennial Symposium*, pp. 239–242 (1994 June).

[32] P. Pestana and J. D. Reiss, "Intelligent Audio Production Strategies Informed by Best Practices," presented at the *AES 53rd International Conference: Semantic Audio* (2014 Jan.), conference paper S2-2.

[33] EBU Tech 3341, "Loudness Metering: 'EBU Mode' Metering to Supplement Loudness Normalisation in Accordance with EBU R128," *European Broadcasting Union* (2016 Jan.).

[34] E. Skovenborg and S. H. Nielsen, "Evaluation of Different Loudness Models with Music and Speech Material," presented at the *117th Convention of the Audio Engineering Society* (2004 Oct.), convention paper 6234.

[35] V. Välimäki et al., "Fifty Years of Artificial Reverberation," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, pp. 1421–1448 (2012 July). https://doi.org/10.1109/TASL.2012.2189567

[36] B. De Man et al., "Perceptual Evaluation of Music Mixing Practices," presented at the *138th Convention of the Audio Engineering Society* (2015 May), convention paper 9235.

[37] B. De Man et al., "The Open Multitrack Testbed," presented at the *137th Convention of the Audio Engineering Society* (2014 Oct.), eBrief 165.

[38] N. Jillings et al., "Web Audio Evaluation Tool: A Browser-Based Listening Test Environment," *12th SMC Conf.* (2015 July).

[39] B. De Man and J. D. Reiss, "APE: Audio Perceptual Evaluation Toolbox for MATLAB," presented at the *136th*

Convention of the Audio Engineering Society (2014 Apr.), eBrief 151.

[40] Recommendation ITU-R BS.1770-4, "Algorithms to Measure Audio Programme Loudness and True-Peak Audio Level" (2015 Oct.).

[41] B. Leonard et al., "The Effect of Playback System on Reverberation Level Preference," presented at the *134th Convention of the Audio Engineering Society* (2013 May), convention paper 8886.

[42] F. E. Toole, "Listening Tests—Turning Opinion into Fact," *J. Audio Eng. Soc.*, vol. 30, pp. 431–445 (1982 June).

[43] R. Schatz et al. "The Impact of Test Duration on User Fatigue and Reliability of Subjective Quality Ratings," *J. Audio Eng. Soc.*, vol. 60, pp. 63–73 (2012 Jan./Feb.).

[44] J. Paulus et al., "Perceived Level of Late Reverberation in Speech and Music," presented at the *130th Convention of the Audio Engineering Society* (2011 May), convention paper 8331.

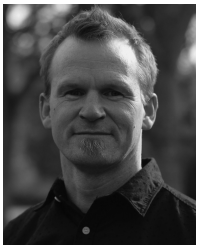
[45] P. Zahorik, "Direct-to-Reverberant Energy Ratio Sensitivity," *J. Acoust. Soc. Amer.*, vol. 112, pp. 2110–2117 (2002 Nov.). https://doi.org/10.1121/1.1506692

[46] H. Herlufsen, "Dual Channel FFT Analysis (Part I)," Brüel & Kjær Technical Review (1984).

[47] D. H. Griesinger, "Quantifying Musical Acoustics through Audibility," *J. Acoust. Soc. Amer.*, vol. 94, p. 1891 (1993 Sep.). https://doi.org/10.1121/1.407513

THE AUTHORS

Brecht De Man



Kirk McNally

Joshua D. Reiss

Brecht De Man is a postdoctoral researcher at the Centre for Digital Music at Queen Mary University of London. Over the course of his Ph.D. at the same institution he has published and presented research on the perception of recording and mix engineering, intelligent audio effects, and the analysis of music production practices. He received an M.Sc. in electronic engineering from the University of Ghent, Belgium, in 2012. An active member of the Audio Engineering Society, he is Vice Chair on the Education Committee, Chair of the London UK Student Section, committee member of the British Section of the AES, and former Chair of the Student Delegate Assembly. In 2013, and again in 2014, he received the HARMAN Scholarship from the AES Educational Foundation. Since 2014, Brecht has been working closely with Yamaha Corporation on the topic of semantic mixing.

•

Kirk McNally is an Assistant Professor of music technology at the University of Victoria in the School of Music. He received his Master of Music degree in sound recording from McGill University. As a recording engineer he has worked with artists including R.E.M, Bryan Adams, Nine Inch Nails, Bad Company, Sloan, The Boston Symphony Orchestra, and the National Youth Orchestra of Canada. Kirk is the program advisor for the undergraduate program in music and computer science as well as the new graduate program in music technology at the University of Victoria. His research interests include sound recording pedagogy, audio archives, and popular music production.

•

Josh Reiss is a Reader with Queen Mary University of London's Centre for Digital Music, where he leads the audio engineering research team. He has investigated music retrieval systems, time scaling and pitch shifting techniques, polyphonic music transcription, loudspeaker design, automatic mixing, sound synthesis, and digital audio effects. His primary focus of research, which ties together many of the above topics, is on the use of state-of-the-art signal processing techniques for professional sound engineering. Dr. Reiss has published over 160 scientific papers, including more than 70 AES publications. His co-authored publications, "Loudness Measurement of Multitrack Audio Content Using Modifications of ITU-R BS.1770" and "Physically Derived Synthesis Model of an Aeolian Tone," were recipients of the 134th AES Convention's Best Peer-Reviewed Paper Award and the 141st AES Convention's Best Student Paper Award, respectively. He co-authored the textbook *Audio Effects: Theory, Implementation and Application*. He is co-founder of the start-up company LandR, providing intelligent tools for audio production. He is a former governor of the AES and was General Chair of the 128th, Program Chair of the 130th, and co-Program Chair of the 138th AES Conventions.