PREDICTING THE PERCEIVED LEVEL OF LATE REVERBERATION USING COMPUTATIONAL MODELS OF LOUDNESS

Christian Uhle, Jouni Paulus

Fraunhofer Institute for Integrated Circuits,
Erlangen, Germany

Jürgen Herre

International Audio Laboratories,
Erlangen, Germany

ABSTRACT

It is well known that the perceived level of reverberation depends on both the input audio signal and the impulse response. This work aims at quantifying this observation and predicting the perceived level of late reverberation based on separate signal paths of direct and reverberant signals, as they appear in digital audio effects. A basic approach to the problem is developed and subsequently extended by considering the impact of the reverberation time on the prediction result. This leads to a linear regression model with two input variables which is able to predict the perceived level with high accuracy, as shown on experimental data derived from listening tests. Variations of this model with different degrees of sophistication and computational complexity are compared regarding their accuracy. Applications include the control of digital audio effects for automatic mixing of audio signals.

***Index Terms*—** Intelligent Digital Audio Effects, Auditory Perception, Artificial Reverberation

1. INTRODUCTION

The aim of this work is to devise a method for predicting the perceived level of reverberation in speech and music when the direct signal and the reverberation impulse response (RIR) are separately available. This functionality is, e.g., desired for applications where an artificial reverberation processor is operated in an automated way and needs to adapt its parameters to the input signal such that the perceived level of the reverberation matches a target value. It is noted that the term *reverberance* while alluding to the same theme, does not appear to have a commonly accepted definition which makes it difficult to use as a quantitative measure in a listening test and prediction scenario.

Artificial reverberation processors are often implemented as linear time-invariant systems and operated in a send-return signal path, as depicted in Figure 1, with pre-delay d, RIR and a scaling factor g for controlling the direct-to-reverberation ratio (DRR). When implemented as parametric reverberation processors, they feature a variety of parameters, e.g. for controlling the shape and the density of the RIR, and the inter-channel coherence (ICC) of the RIRs for multi-channel processors in one or more frequency bands.

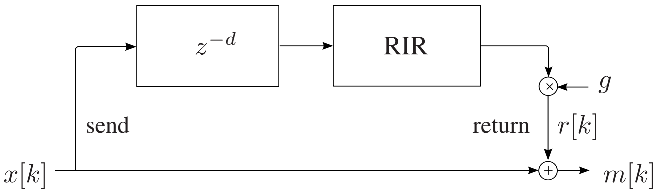


Fig. 1. Block diagram of an artificial reverberation processor.

These parameters have an impact on the resulting audio signal in terms of perceived level, distance, room size, coloration and sound quality. Furthermore, the perceived characteristics of the reverberation depend on the temporal and spectral characteristics of the input signal [1]. Focusing on a very important sensation, namely loudness, it can be observed that the loudness of the perceived reverberation is monotonically related to the non-stationarity of the input signal. Intuitively speaking, an audio signal with large variations in its envelope excites the reverberation at high levels and allows it to become audible at lower levels. In a typical scenario where the long-term DRR expressed in decibels is positive, the direct signal can mask the reverberation signal almost completely at time instances where its energy envelope increases. On the other hand, whenever the signal ends, the previously excited reverberation tail becomes apparent in gaps exceeding a minimum duration determined by the slope of the post-masking (at maximum 200 ms) and the integration time of the auditory system (at maximum 200 ms for moderate levels).

To illustrate this, Figure 2 shows the time signal envelopes of a synthetic audio signal and of an artificially generated reverberation signal. An RIR with a short pre-delay of 50 ms is used here, omitting early reflections and synthesizing the late part of the reverberation with exponentially decaying white noise [2]. The input signal has been generated from a harmonic wide-band signal and an envelope function such that one event with a short decay and a second event with a long decay are perceived. While the long event produces more total reverberation energy, it comes to no surprise that it is the short sound which is perceived as being more reverber-

978-1-4577-0274-7/11/$26.00 ©2011 IEEE

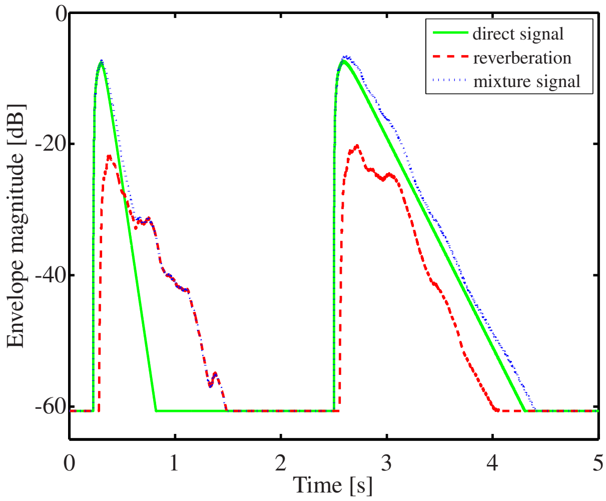


Fig. 2. Example of time signal envelopes of an audio signal (solid line), the reverberation signal (dashed line), and the mixture of both signals (dotted line).

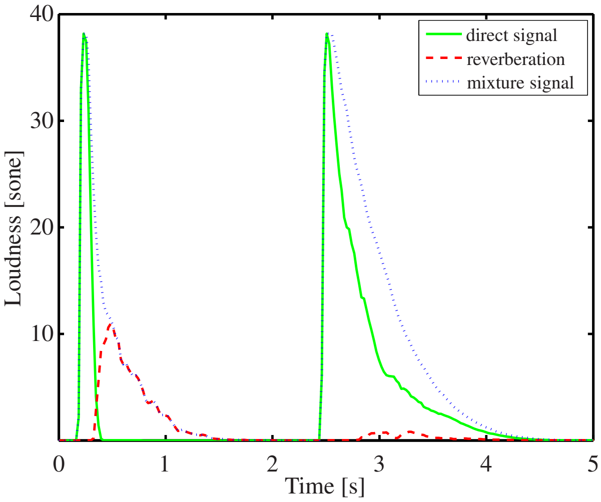


Fig. 3. Loudness and partial loudness of the example signal: total loudness (dotted line), partial loudness of direct signal (solid line) and of the reverberation signal (dashed line).

ant. Where the decaying slope of the longer event masks the reverberation, the short sound already disappeared before the reverberation has built up and thereby a gap is open in which the reverberation is perceived. Please note that the definition of masking used here includes both complete and partial masking [3].

Although such observations have been made many times [4, 5, 6], it is still worth emphasizing them because it illustrates qualitatively why models of partial loudness can be applied in the context of this work. In fact, it has been pointed out that the perception of reverberation arises from stream segregation processes in the auditory system [4, 5, 6] and is influenced by the partial masking of the reverberation due to the direct sound.

The considerations above motivate the use of loudness models. Related investigations were performed by Lee et.al. and focus on the prediction of the subjective decay rate of RIRs when listening to them directly [7] and on the effect of the playback level on reverberance [8]. A predictor for reverberance using loudness-based early decay times is proposed in [9]. In contrast to this work, the prediction methods proposed here process the direct signal and the reverberation signal with a computational model of partial loudness (and with simplified versions of it in the quest for low-complexity implementations) and thereby consider the influence of the input (direct) signal on the sensation. Recently, Tsilfidis and Mourjopoulus [10] investigated the use of a loudness model for the suppression of the late reverberation in single-channel recordings. An estimate of the direct signal is computed from the reverberant input signal using a spectral subtraction method, and a reverberation masking index is derived by means of a computational auditory masking model, which controls the dereverberation processing.

The implementation of the loudness model used here follows the descriptions in [11, 12] with modifications as detailed in Subsection 2.1. The training and the validation of the prediction uses data from listening tests described in [13] and briefly summarized in Subsection 2.2. The application of the loudness model for predicting the perceived level of late reverberation is described in Subsections 2.3 and 2.4. Experimental results are presented in Section 3 and conclusions are given in Section 4.

2. DERIVATION OF THE METHOD

This section describes the implementation of a model of partial loudness, the listening test data that was used as ground truth for the computational prediction of the perceived level of reverberation, and a proposed prediction method which is based on the partial loudness model.

2.1. A model of partial loudness

The loudness model computes the partial loudness $N_{x,n}[k]$ of a signal $x[k]$ when presented simultaneously with a masking signal $n[k]$

$$N_{x,n}[k] = f(x[k], n[k]). \tag{1}$$

Although early models have dealt with the perception of loudness in steady background noise, some work exists on loudness perception in backgrounds of co-modulated random noise [14], complex environmental sounds [12], and music signals [15]. Figure 3 illustrates the total loudness and the partial loudness of its components of the example signal shown in Figure 2, computed with the loudness model used here.

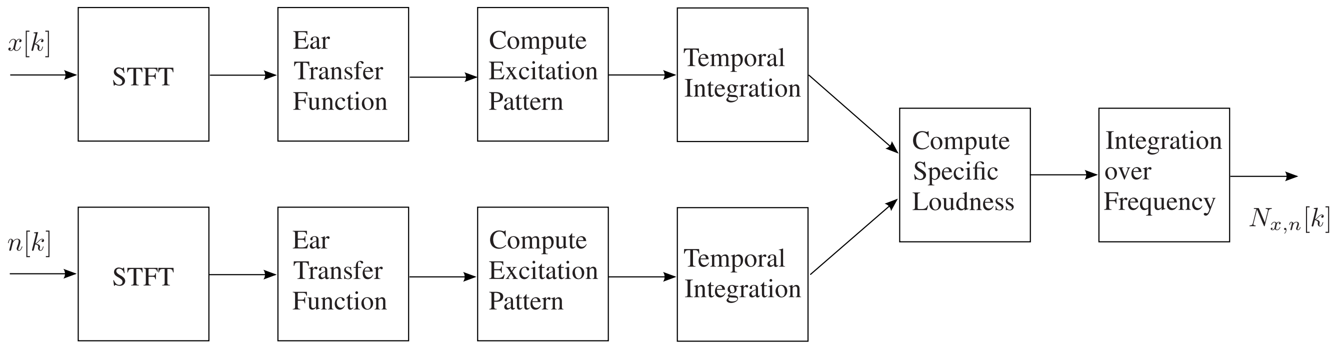


Fig. 4. Block diagram of the loudness model.

The model used in this work is similar to the models in [11, 12] which itself drew on earlier research by Fletcher, Munson, Stevens, and Zwicker, with some modifications as described in the following. A block diagram of the loudness model is shown in Figure 4. The input signals are processed in the frequency domain using a Short-time Fourier transform (STFT). In [12], 6 DFTs of different lengths are used in order to obtain a good match for the frequency resolution and the temporal resolution to that of the human auditory system at all frequencies. In this work, only one DFT length is used for the sake of computational efficiency, with a frame length of 21 ms at a sampling rate of 48 kHz, 50% overlap and a Hann window function. The transfer through the outer and middle ear is simulated with a fixed filter. The excitation function is computed for 40 auditory filter bands spaced on the equivalent rectangular bandwidth (ERB) scale using a level dependent excitation pattern. In addition to the temporal integration due to the windowing of the STFT, a recursive integration is implemented with a time constant of 25 ms, which is only active at times where the excitation signal decays.

The specific partial loudness, i.e., the partial loudness evoked in each of the auditory filter band, is computed from the excitation levels from the signal of interest (the stimulus) and the interfering noise according to Equations (17)-(20) in [11]. These equations cover the four cases where the signal is above the hearing threshold in noise or not, and where the excitation of the mixture signal is less than 100 dB or not. If no interfering signal is fed into the model, i.e. $n\left[k\right] = 0$, the result equals the total loudness $N_x\left[k\right]$ of the stimulus $x\left[k\right]$.

2.2. Description of listening tests and experimental data

In order to assess the suitability of the described loudness model for the task of predicting the perceived level of the late reverberation, a corpus of ground truth generated from listener responses is necessary. To this end, data from an investigation featuring several listening tests [13] is used in this paper which is briefly summarized in the following. Each listening test consisted of multiple graphical user interface screens which presented mixtures of different direct signals with different conditions of artificial reverberation. The listeners were asked to rate the perceived amount of reverberation on a scale from 0 to 100 points. In addition, two anchor signals were presented at 10 points and at 90 points. The anchor signals were created from the same direct signal with different conditions of reverberation.

The direct signals used for creating the test items were monophonic recordings of speech, individual instruments and music of different genres with a length of about 4 seconds each. The majority of the items originated from anechoic recordings but also commercial recordings with a small amount of reverberation were used.

The RIRs represent late reverberation and were generated using exponentially decaying white noise with frequency dependent decay rates. The decay rates are chosen such that the reverberation time decreases from low to high frequencies, starting at a base reverberation time T_{60}. Early reflections were neglected in this work. The reverberation signal $r\left[k\right]$ and the direct signal $x\left[k\right]$ were scaled and added such that the ratio of their average loudness measure according to ITU-R BS.1770 [16] matches a desired DRR and such that all test signal mixtures have equal long-term loudness. All participants in the tests were working in the field of audio and had experience with subjective listening tests.

The ground truth data used for the training and the verification / testing of the prediction method were taken from two listening tests and are denoted by A and B, respectively. The data set A consisted of ratings of 14 listeners for 54 signals. The listeners repeated the test once and the mean rating was obtained from all of the 28 ratings for each item. The 54 signals were generated by combining 6 different direct signals and 9 stereophonic reverberation conditions, with $T_{60} \in \{1, 1.6, 2.4\}$ s and $DRR \in \{3, 7.5, 12\}$ dB, and no pre-delay.

The data in B were obtained from ratings of 14 listeners for 60 signals. The signals were generated using 15 direct signals and 36 reverberation conditions. The reverberation conditions sampled four parameters, namely T_{60}, DRR, pre-delay, and ICC. For each direct signal 4 RIRs were chosen such that two had no pre-delay and two had a short pre-delay

	r	MAE	$RMSE$
$\hat{R}_b$, T	0.76	9.7	12.1
$\hat{R}_b$, V	0.76	10.4	12.6
$\hat{R}_e$, T	0.85	8.3	10.0
$\hat{R}_e$, V	0.85	8.3	10.6

Table 1. Evaluation metrics of the prediction of $\hat{R}_b$ and $\hat{R}_e$ for training (T) and testing (V).

of 50 ms, and two were monophonic and two were stereophonic.

2.3. Using the loudness model for predicting the perceived level of late reverberation

The basic input feature for the prediction method is computed from the difference of the partial loudness $N_{r,x}\left[k\right]$ of the reverberation signal $r\left[k\right]$ (with the direct signal $x\left[k\right]$ being the interferer) and the loudness $N_{x,r}\left[k\right]$ of $x\left[k\right]$ (where $r\left[k\right]$ is the interferer), according to Equation 2.

$$\Delta N_{r,x}\left[k\right] = N_{r,x}\left[k\right] - N_{x,r}\left[k\right] \qquad (2)$$

The rationale behind Equation (2) is that the difference $\Delta N_{r,x}\left[k\right]$ is a measure of how strong the sensation of the reverberation is compared to the sensation of the direct signal. Taking the difference was also found to make the prediction result approximately invariant with respect to the playback level. The playback level has an impact on the investigated sensation [17, 8], but to a more subtle extent than reflected by the increase of the partial loudness $N_{r,x}$ with increasing playback level. Typically, musical recordings sound more reverberant at moderate to high levels (starting at about 75-80 dB SPL) than at about 12 to 20 dB lower levels. This effect is especially obvious in cases where the DRR is positive, which is valid “for nearly all *recorded* music” [18], but not in all cases for concert music where “listeners are often well beyond the critical distance” [6].

The decrease of the perceived level of the reverberation with decreasing playback level is best explained by the fact that the dynamic range of reverberation is smaller than that of the direct sounds (or, a time-frequency representation of reverberation is more dense whereas a time-frequency representation of direct sounds is more sparse [19]). In such a scenario, the reverberation signal is more likely to fall below the threshold of hearing than the direct sounds do.

2.4. Prediction model

The prediction methods described in the following are linear and use a least squares fit for the computation of the model coefficients. The simple structure of the predictor is advantageous in situations where the size of the data sets is limited, which could lead to overfitting of the model when using regression methods with more degrees of freedom, e.g. neural networks. The baseline predictor $\hat{R}_b$ is derived by the linear regression according to Equation (3) with coefficients a_i, with K being the length of the signal in frames,

$$\hat{R}_b = a_0 + a_1 \frac{1}{K} \sum_{k=1}^{K} \Delta N_{r,x}\left[k\right]. \qquad (3)$$

The model has only one independent variable, i.e. the mean of $\Delta N_{r,x}\left[k\right]$. To track changes and to be able to implement a real-time processing, the computation of the mean can be approximated using a leaky integrator, but this is not investigated here. The model parameters derived when using data set A for the training are $a_0 = 48.2$ and $a_1 = 14.0$, where a_0 equals the mean rating for all listeners and items.

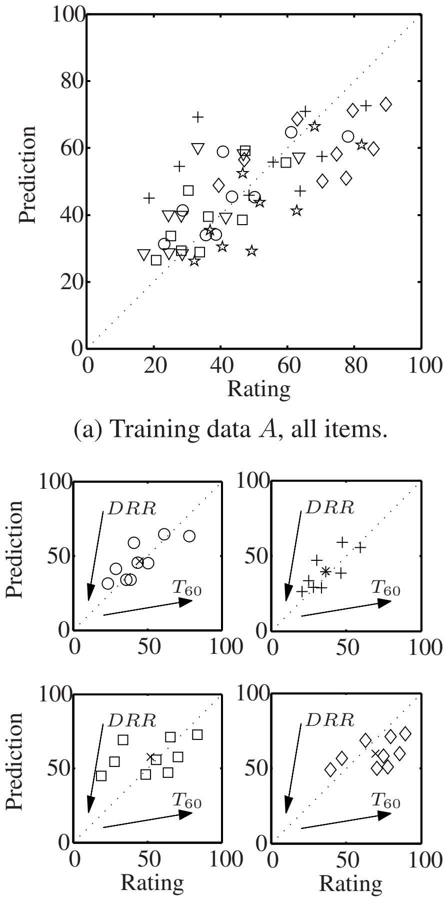
Figure 5 depicts the predicted sensations for data set A. It can be seen that the predictions are moderately correlated with the mean listener ratings with a correlation coefficient of 0.71. Please note that the choice of the regression coefficients does not affect this correlation. As shown in the lower plot, for each mixture generated by the same direct signals, the points exhibit a characteristic shape centered close to the diagonal. This shape indicates that although the baseline model $\hat{R}_b$ is able to predict R to some degree, it does not reflect the influence of T_{60} on the ratings. The visual inspection of the data points suggests a linear dependency on T_{60}. If the value of T_{60} is known, as is the case when controlling an audio effect, it can be easily incorporated into the linear regression model to derive an enhanced prediction

$$\hat{R}_e = a_0 + a_1 \frac{1}{K} \sum_{k=1}^{K} \Delta N_{r,x}\left[k\right] + a_2 T_{60}. \qquad (4)$$

The model parameters derived from the data set A are $a_0 = 48.2$, $a_1 = 12.9$, $a_2 = 10.2$. The results are shown in

	r	MAE	$RMSE$
ΔN_{m-x}, T	0.81	9.0	11.2
ΔN_{m-x}, V	0.81	9.0	11.1
ΔN_{r-m}, T	0.80	9.5	11.3
ΔN_{r-m}, V	0.80	10.1	12.7
ΔN_{r-x}, T	0.83	8.8	10.5
ΔN_{r-x}, V	0.83	9.1	11.1

Table 2. Evaluation metrics obtained when using the total loudness of separated signals and mixture signal, for training (T) and testing (V). See Equations (5)-(7) for an explanation of the loudness features.



(a) Training data A, all items.

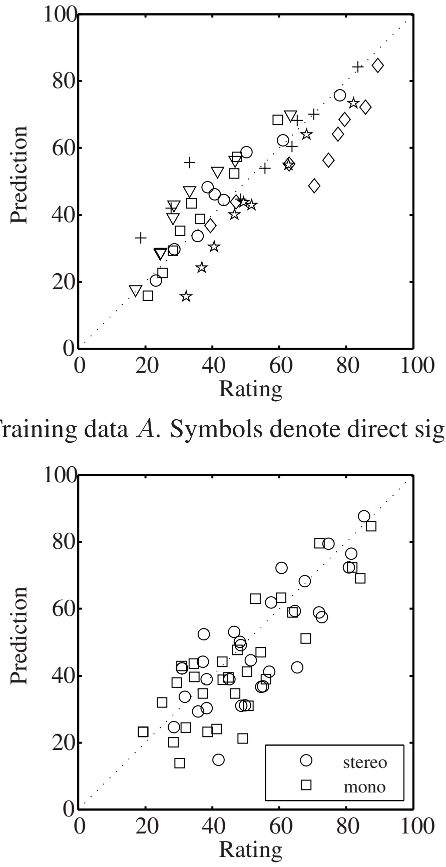
(b) A subset of the training data.

Fig. 5. The predictions of the baseline model and the references for data set A. Symbols denote mixtures originating from the same direct signal. The correlation coefficient is 0.71. The lower plot shows all mixtures originating from the first 4 direct signals separately, together with the centroids of their points (x-mark). The arrows show the direction of change of the RIR parameters.

Figure 6 separately for each of the data sets. The evaluation of the results is described in more detail in the next section.

3. FURTHER EXPERIMENTS AND RESULTS

In the following, the models are evaluated using the correlation coefficient r, the mean absolute error (MAE) and the root mean squared error ($RMSE$) between the mean listener ratings and the predicted sensation. The experiments are performed as two-fold cross-validation, i.e. the predictor is trained with data set A and tested with data set B, and the experiment is repeated with B for training and A for testing. The evaluation metrics obtained from both runs are averaged, separately for the training and the testing.



(a) Training data A. Symbols denote direct signals.

(b) Test data B. Symbols denote mono and stereo RIRs.

Fig. 6. The listener ratings and the outputs of the prediction model $\hat{R}_e$, trained with data set A and tested with data set B.

The results are shown in Table 1 for the prediction models $\hat{R}_b$ and $\hat{R}_e$. The predictor $\hat{R}_e$ yields accurate results with an $RMSE$ of 10.6 points. The average of the standard deviation of the individual listener ratings per item are given as a measure for the dispersion from the mean (of the ratings of all listeners per item) as $\overline{\sigma}_A = 13.4$ for data set A and $\overline{\sigma}_B = 13.6$ for data set B. The comparison to the $RMSE$ indicates that $\hat{R}_e$ is at least as accurate as the average listener in the listening test.

The accuracies of the predictions for the data sets differ slightly, e.g. for $\hat{R}_e$ both MAE and $RMSE$ are approximately one point below the mean value (as listed in the table) when testing with data set A and one point above average when testing with data set B. The fact that the evaluation metrics for training and test are comparable indicates that overfitting of the predictor has been avoided.

In order to facilitate an economic implementation of such prediction models, the following experiments investigate how the use of loudness features with less computational complexity influence the precision of the prediction result. The experiments focus on replacing the partial loudness computation by estimates of total loudness and on simplified implementations of the excitation pattern.

Instead of using the partial loudness difference $\Delta N_{r,x}[k]$, three differences of total loudness estimates are examined, with the loudness of the direct signal $N_x[k]$, the loudness of the reverberation $N_r[k]$, and the loudness of the mixture signal $N_m[k]$, as shown in Equations (5)-(7), respectively.

$$\Delta N_{m-x}[k] = N_m[k] - N_x[k] \tag{5}$$

Equation (5) is based on the assumption that the perceived level of the reverberation signal can be expressed as the difference (increase) in overall loudness which is caused by adding the reverb to the dry signal.

Following a similar rationale as for the partial loudness difference in Equation (2), loudness features using the differences of total loudness of the reverberation signal and the mixture signal or the direct signal, respectively, are defined in Equations (6) and (7). The measure for predicting the sensation is derived from as the loudness of the reverberation signal when listened to separately, with subtractive terms for modelling the partial masking and for normalization with respect to playback level derived from the mixture signal or the direct signal, respectively.

$$\Delta N_{r-m}[k] = N_r[k] - N_m[k] \tag{6}$$

$$\Delta N_{r-x}[k] = N_r[k] - N_x[k] \tag{7}$$

Table 2 shows the results obtained with the features based on the total loudness and reveals that in fact two of them, $\Delta N_{m-x}[k]$ and $\Delta N_{r-x}[k]$, yield predictions with nearly the same accuracy as $\hat{R}_e$. It remains to be investigated whether this unexpectedly good performance generalizes and also holds for more diverse reverberation conditions, especially when using larger values for the pre-delay.

Finally, in an additional experiment, the influence of the implementation of the spreading function is investigated. This is of particular significance for many application scenarios, because the use of the level dependent excitation patterns demands implementations of high computational complexity. The experiments with a similar processing as for $\hat{R}_e$ but using one loudness model without spreading and one loudness model with level-invariant spreading function led to the results shown in Table 3. Somewhat surprisingly, the influence of the spreading seems to be negligible. It is assumed that for spectrally sparse signals the spreading will play a larger role, however, this has not been demonstrated with the data sets used here.

	r	MAE	$RMSE$
no spreading, T	0.84	8.4	10.1
no spreading, V	0.84	8.8	10.6
fixed spread., T	0.85	8.3	10.1
fixed spread., V	0.85	8.7	10.7

Table 3. Evaluation metrics for loudness features without spreading and with level-invariant spreading, for training (T) and testing (V).

4. CONCLUSIONS AND FUTURE WORK

This work presented an investigation in simple and robust prediction of the perceived level of late reverberation in speech and music using loudness models of varying computational complexity. The prediction models have been trained and evaluated using subjective data derived from three listening tests. As a starting point, the use of a partial loudness model has led to a prediction model with high accuracy when the T_{60} of the RIR is known. This result is also interesting from the perceptual point of view when considering that model of partial loudness has not been developed with stimuli of direct and reverberant sound.

Subsequent modifications of the computation of the input features for the prediction method led to a series of simplified models which were shown to achieve comparable performance for the data sets at hand. These modifications included the use of total loudness models and simplified spreading functions. Future work will investigate the prediction of sensations evoked by more diverse RIRs including early reflections and larger pre-delays. Also, the perceived loudness contribution of other types of additive audio effects can be investigated in a similar manner.

5. REFERENCES

[1] A. Czyzewski, "A method for artificial reverberation quality testing," *J. Audio Eng. Soc.*, vol. 38, pp. 129–141, 1990.

[2] J.A. Moorer, "About this reverberation business," *Computer Music Journal*, vol. 3, 1979.

[3] B. Scharf, "Fundamentals of auditory masking," *Audiology*, vol. 10, pp. 30–40, 1971.

[4] W. G. Gardner and D. Griesinger, "Reverberation level matching experiments," in *Proc. of the Sabine Centennial Symposium, Acoust. Soc. of Am.*, 1994.

[5] D. Griesinger, "How loud is my reverberation," in *Proc. of the AES 98th Conv.*, 1995.

[6] D. Griesinger, "Further investigation into the loudness of running reverberation," in *Proc. of the Institute of Acoustics (UK) Conference*, 1995.

[7] D. Lee and D. Cabrera, "Effect of listening level and background noise on the subjective decay rate of room impulse responses: Using time-varying loudness to model reverberance," *Applied Acoustics*, vol. 71, pp. 801–811, 2010.

[8] D. Lee, D. Cabrera, and W.L. Martens, "Equal reverberance matching of music," in *Proc. of Acoustics*, 2009.

[9] D. Lee, D. Cabrera, and W.L. Martens, "Equal reverberance matching of running musical stimuli having various reverberation times and SPLs," in *Proc. of the 20th International Congress on Acoustics*, 2010.

[10] A. Tsilfidis and J. Mourjopoulus, "Blind single-channel suppression of late reverberation based on perceptual reverberation modeling," *J. Acoust. Soc. Am*, vol. 129, pp. 1439–1451, 2011.

[11] B.C.J. Moore, B.R. Glasberg, and T. Baer, "A model for the prediction of thesholds, loudness, and partial loudness," *J. Audio Eng. Soc.*, vol. 45, pp. 224–240, 1997.

[12] B.R. Glasberg and B.C.J. Moore, "Development and evaluation of a model for predicting the audibility of time-varying sounds in the presence of background sounds," *J. Audio Eng. Soc.*, vol. 53, pp. 906–918, 2005.

[13] J. Paulus, C. Uhle, and J. Herre, "Perceived level of late reverberation in speech and music," in *Proc. of the AES 130th Conv.*, 2011.

[14] J.L. Verhey and S.J. Heise, "Einfluss der Zeitstruktur des Hintergrundes auf die Tonhaltigkeit und Lautheit des tonalen Vordergrundes (in German)," in *Proc. of DAGA*, 2010.

[15] C. Bradter and K. Hobohm, "Loudness calculation for individual acoustical objects within complex temporally variable sounds," in *Proc. of the AES 124th Conv.*, 2008.

[16] International Telecommunication Union, Radiocomunication Assembly, "Algorithms to measure audio programme loudness and true-peak audio level," Recommendation ITU-R BS.1770, 2006, Geneva, Switzerland.

[17] S. Hase, A. Takatsu, S. Sato, H. Sakai, and Y. Ando, "Reverberance of an existing hall in rrelation to both subsequent reverberation time and SPL," *J. Sound Vib.*, vol. 232, pp. 149–155, 2000.

[18] D. Griesinger, "The importance of the direct to reverberant ratio in the perception of distance, localization, clarity, and envelopment," in *Proc. of the AES 126th Conv.*, 2009.

[19] C. Uhle, A. Walther, O. Hellmuth, and J. Herre, "Ambience separation from mono recordings using Nonnegative Matrix Factorization," in *Proc. of the AES 30th Conf.*, 2007.