Acoust. Sci. & Tech. **41**, 4 (2020) ©2020 The Acoustical Society of Japan

PAPER

An auditory scaling method for reverb synthesis from a single two-dimensional image

Homare Kon* and Hideki Koike

School of Computing, Tokyo Institute of Technology,
2–12–1 Ookayama, Meguro-ku, Tokyo, 152–8550 Japan

(*Received 17 July 2019, Accepted for publication 3 January 2020*)

Abstract: We present an auditory scaling method to generate reverberant sounds that more appropriately match the expected auditory impression of a space in a 2D image. Since the conventional method uses linear scale for the regression parameters of reverberation characteristics, correspondence with the human sense scale has been not considered. We have incorporated concepts from psychoacoustics into the reverberate parameters to improve regression performance in an actual environment, including the sound-masking effect, equal-loudness curves, and subjectively-equal reverberation time. Estimation errors in our scaling method were significantly lower than in comparison with previously presented results. The proposed reverb synthesis method was then evaluated in tests, using several scenes to demonstrate its benefits. Our reverb synthesis method can reproduce plausible reverberant sounds from 2D images, which can be used in mixed and augmented-reality applications.

Keywords: Reverberation, Regression, Place recognition, Image to sound

PACS number: 43.60.−c [doi:10.1250/ast.41.675]

1. INTRODUCTION

To reproduce immersive experiences in mixed-reality (MR) or augmented-reality (AR) modes, virtual objects must be realistically presented [1]. The sense of immersion when interacting with virtual objects depends on audio-visual information that matches users' expectations [2]. Compared with interactive content, such as virtual reality or video games, the environment through which a user will move in an MR/AR application is difficult to anticipate. As sound differently reflects in different spaces, and users are able to notice when a room sounds unnatural [3,4]. The MR/AR tool must subsequently replicate the reverberation characteristics of a room to match users' expected characteristics.

A physical approach is common, by which the AR application simulates the physical phenomenon of reverberation. In these simulations, the transfer path and the reflectance path are calculated geometrically, using 3D reconstruction from multiple images [5,6]. Some researchers have tried to simulate actual physical spaces via waveform levels by considering the reflectance characteristics of estimated materials in the walls of such spaces [7].

*e-mail: kon.h.aa@m.titech.ac.jp

Perceptual approaches can also be used, which focus less on the physics of the room and more on the user's perspective. Kon and Koike, for instance, focus on the ability of expert audio engineers to estimate reverberation characteristics from images, subsequently demonstrating the feasibility of training convolutional neural networks (CNNs) in this expert skill [8,9]. Training an automated system to behave as an audio expert facilitates the synthesis of reverb without the need for a 3D model of the space.

However, the reproduction of reverberation for MR/AR applications remains somewhat difficult. Physical approaches need several images to reconstruct a 3D model of the space. Users therefore need to explore and observe a room from several perspectives, so that the reverb module can prepare a 3D model, before generating synthetic reverb. The amount of computation for these calculations is proportional to the size of the room, which can be extremely large [7]. As for the perceptual approach, large numerical errors remain in the estimation of reverberation characteristics [8,9]. Moreover, this approach does not mention how subjectively those errors affect impressions on audio-visual information matching.

One advantage of the perceptual approach is that the computational cost is constant, regardless of the spatial

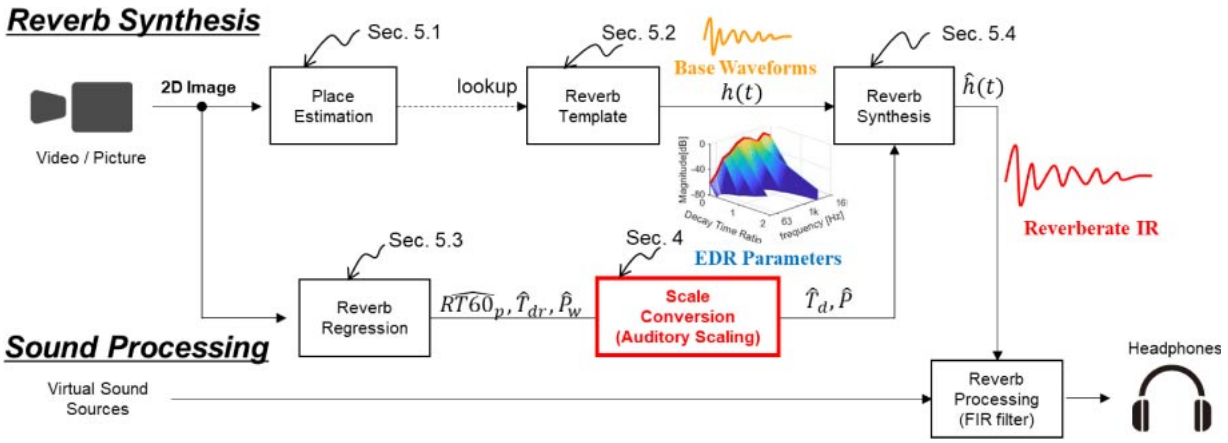


Fig. 1 System overview for reverb synthesis from a 2D image.

extent of the environment in question. Therefore, in this paper, we propose a method to improve the estimation of reverberation characteristics via the perceptual approach, and evaluate our method in user tests.

System Overview using Proposed Method

Figure 1 presents a block diagram, which is the system overview in this paper. Our reverberation synthesis system is an extended system based on the method for synthesizing a room impulse response (IR) suggested by Moorer [10,11]. In preliminary research for the design of the overall architecture as the perceptual approach, we interviewed professional and expert acoustic engineers about how they accomplish the feat of estimating the acoustic characteristics of a room using only images. Their comments were as follows:

- Broad characteristics are determined according to visual cues about the context of the space.
- Characteristics, such as timbre and reverberation time, are imagined from memories of similar spaces.

Therefore, we assumed four steps for a reverberation synthesis system from a 2D image. First the recognition step recognizes the scene, after which recall step associates a broad set of reverberation characteristics with the recognized scene. This is followed by the imagination phase, which predicts reverberation characteristics based on the differences between the impression of the current image and the images associated with known reverb characteristics. Finally, the rendering phase synthesizes the reverberation. In other word, the recall step corresponds to the Reverb Template block, which refers base waveforms based on the result of the Place Estimation block, which corresponds to the recognition step. Then, the Reverb Regression block, or the imagination phase, estimates reverberation characteristics and synthesizes the reverberation IR. Finally, the Reverb Processing block, or the rendering phase creates the reverberation sound.

In the conventional estimation method [8,9], the output parameters of the reverb regression block are linearly weighted. However, we considered that perceptual inaccuracies can be reduced if weighted inference, which is suitable for human senses can be made. For example, since the discrimination thresholds for reverberations are tested by using rate of change of the reference stimulus, the conventional linear scaling of reverberation time is considered inappropriate [12,13]. To improve the estimation of reverberation characteristics, we apply auditory scaling techniques based on hearing characteristics, assuming that an AR/MR system is used in an actual environment. For example, we apply psychoacoustic concepts, such as sound-masking effects [14,15] and the quantization of human perception, which is known as the Weber-Fechner law in the field of psychophysics [16]. Therefore, the output parameters of the reverb regression block are scaled with our auditory scaling method, and scaled output parameters are converted to linear scale in the Scale Conversion block.

In Sect. 3, we describe the reverberation IR and image database for the reverberation characteristics estimation. Section 4 introduces auditory scaling techniques that improve regression accuracy. We then describe the detailed implementation of our system in Sect. 5. In Sects. 6 and 7, we exhibit the results obtained by our system and user evaluation. Finally, Sect. 8 summarize the study.

2. RELATED WORK

This section reviews the literature on visual image recognition and the reproduction of reverberation.

2.1. Reverberation

Interactive sound rendering in MR/AR comprises three types of layers: modeling, propagation, and 3D audio rendering. The present study relates to reverberation in 3D audio rendering [5]. As artificial reverberation has been studied for more than 50 years, many synthesis methods are

now available [17]. Both physical and perceptual approaches can be used in the reproduction of reverberation characteristics [18]. Physical approaches involve simulating the reflection of infinite sounds in a certain space at the level of waveforms. Perceptual approaches use findings in the psychology of perception to reproduce reverberation, using a probabilistic model.

The perceptual approach also employs a digital filter to avoid any discomfort based on measured reverberation characteristics. In general, auditory approaches can reproduce natural-seeming reverberation with lower computational cost, but they require known reverberation characteristics. For example, for the reverberation time measured $RT60$ by energy decay curve (EDC) [19], the decay time for each room frequency and the energy decay relief (EDR) [2,20] are required as parameters, to calculate reverberation characteristics from a perceptual perspective. Physical approaches simulate an IR using boundary conditions and a 3D model of the room, and sound sources are convolved with simulated IR as the coefficients of a finite IR (FIR) filter [21]. Both numerical simulations [22] and geometric models [23] require 3D models in addition to boundary conditions such as restitution coefficients. In previously reported reverberation-reproduction methods, the targeted characteristics, 3D models, or boundary conditions must be known in advance.

Several methods for estimating reverberation characteristics from 2D images have been introduced over the recent years. Schissler *et al.*, for instance, propose a novel method for boundary-condition problems that reconstructs a 3D model from multiple images and subsequently estimates the boundary conditions from the images, using the materials in context (MINC) database [7,24]. Alternatively, as a novel perceptual approach, CNNs have been employed to estimate reverberation characteristics from 2D images of a space [8,9]. This perceptual approach has the advantage of lower computational cost, since there is no need for 3D reconstruction compared with the physical approach. However, it is limited by problems of estimation scale, which does not consider the extents of human sense. This means the estimation error is still large.

2.2. Image Recognition

Since AlexNet [25] was presented, CNNs have been used to improve deep-learning performance in the field of image recognition [26,27]. In this study, we apply CNNs for scene recognition to identify the broad reverberation characteristics of a room. The datasets for scene recognition are Places 109, Places 365 [28], SUN 397, and SUN 908 [29]. The number of classifications varies depending on the data set. Although the datasets are different, they have all been successfully used to train various neural networks [27,30,31].

Therefore, image recognition with datasets for scene recognition can be useful to determine broad reverberation characteristics according to visual cues about the places in this paper.

3. REVERBERATION AND IMAGE DATABASE

We created a reverberation database for reverberation characteristics estimation. This database contains various reverberate IRs of outdoor and indoor spaces, which have been taken from the libraries of DAW [32–34], the OpenAIR library [35], and IRs that we measured at several locations. The databese contains a total of 523 scenes and 5106 IRs. To reduce variation in the data due to equipment, we excluded data for which the recording equipment was unknown or those not measured with a loudspeaker. For example, sounds of a balloon-pop, pistols, or hand claps were not included in the database.

The number of images is approximately 60,000. Images were listed in the manual, together with the measurement description given in those libraries. Since some scenes have panoramic images [33], they were cut out at roughly a 55-degree viewing angle.

We manually classified IRs based on metadata, using the Level 2 hierarchy regulated by SUN [29]. We chose this level of classification detail over a more fine-grained scheme because detailed categories of the SUN and Places365 databases may not have returned enough IRs for the tests described below. For example, the fine-grained categories of “snowfield,” “ski_resort,” and “ski_slope” in Places365 are expected to have the same reverberation characteristics. The inclusion of all these similar categories in the database would mean including too many categories to efficiently classify reverberation characteristics. In addition, “Level 2 Outdoor, Natural Man-made” was excluded from the classification for the same reason. For example, since the category “tree_house” is an object visible in outdoor nature, it is unlikely to affect the reverberation of sounds propagating the outdoor scene. As mentioned above, we labeled 15 categories in total, which are listed in Table 1.

4. AUDITORY SCALING

This section discusses the auditory scaling of reverberation characteristics, assuming a real environment and using the auditory-perception approach.

4.1. Definition of the Reverberation Parameter

We apply the EDR as reverberation parameters in the same manner as [8]. To summarize this step, $h(\tau)$ is defined as the measured reverberate IR:

Table 1 IR classification distribution with Level 2 hierarchy specified by [29].

Level 1 Category Name / - Level 2 Category Name	Samples
Indoor /	
- Shopping and dining	257
- workplace	130
- home or hotel	221
- transportation	341
- sports and leisure	515
- cultural	2,883
Outdoor, Natural /	
- water, ice, snow	64
- mountains, hills, desert, sky	79
- forest, field, jungle	125
Outdoor, Man-made /	
- transportation	40
- cultural or historical building place	233
- sports fields, parks, leisure, spaces	96
- industrial and construction	30
- houses, cabin, gardens, and farms	32
- commercial buildings, shops, markets, etc.	60

$$EDR(t,\omega) = 10 \cdot \log_{10}\left|\int_t^{\infty} h(\tau)\cdot e^{-j\omega\tau}d\tau\right|^2 \quad (1)$$

where ω is the angular frequency.

In this paper, the frequency bandwidth is a whole-octave band, with nominal frequencies from 31.5 Hz to 16 kHz. According to [2], the EDR is fully characterized by the frequency dependance of the initial power spectrum $P(f_k)$ and decay time $T_d(f_k)$, where k is the band number in a whole-octave band. The decay time is calculated for reverberation characteristics with a linear curve fitting for each frequency band [2,36], which is replaced by the slope of the attenuation in each frequency band. In addition to these factors, the overall reverberation time $RT60$ is applied as the reverberation parameter [19].

4.2. Reverberation Time Scaling

We convert $RT60$ into a perceptual quantity based on the integral Weber-Fechner law [16,37].

$$p = K \cdot \ln\frac{S}{S_0} \quad (2)$$

where p is the humanly perceived stimulus intensity, the constant K is sense-specific to the sense modality, S is the stimulus, and S_0 is the perceptible threshold. According to this law, the subjective sensation is proportional to the logarithm of the stimulus intensity. In this paper, S_0 is assumed to be 1, $K = 1$, and $RT60p$ is the sensory stimulus of reverberation time, which can then be denoted as

$$RT60p = \ln(RT60). \quad (3)$$

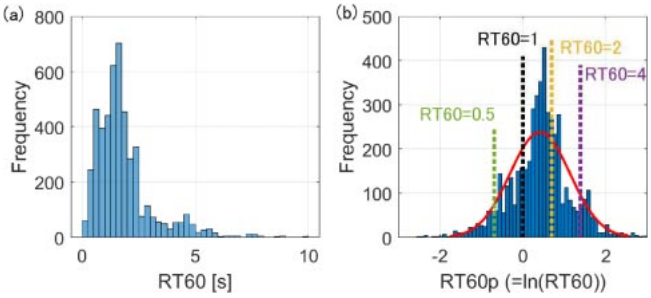


Fig. 2 Histograms of reverberation time of 4,857 samples. (a) is displayed in linear scale of $RT60$. (b) is displayed in logarithmic scale. The red line marks the normal distribution.

Figure 2 shows a histogram of the distribution of reverberation times. We analyzed reverberation times ranging from 0.08 s to 36.75 s. The distribution of $RT60$ tends to have fewer samples as the reverberation time gets longer, that of $RT60p$ is close to the normal distribution. In terms of calculating the mean and median, the median value of $RT60$ was 1.58 s, the mean value of $RT60$ was 1.98 s, and the mean value of $RT60p$ was 1.51 s on the linear scale. The mean value of $RT60$ is increased by the long reverberation times. The mean value of $RT60p$ is close to the median value of $RT60$ on the linear scale.

4.3. Initial Power Scaling

The supposed audible range of a practical AR/MR environment is depicted in Fig. 3, which considers the threshold of pain [38], an equal-loudness contour [39], and sound masking effects [14,15] in a daily life environment.

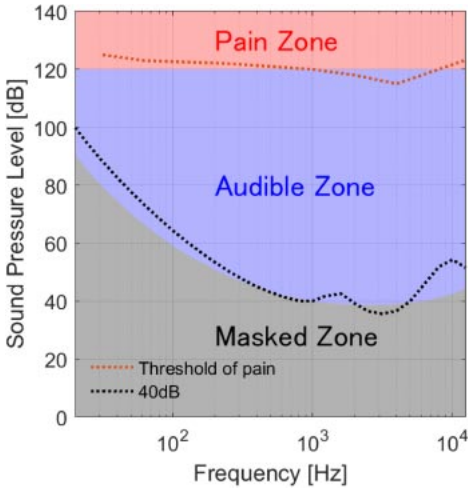


Fig. 3 Examples of equal-loudness contours specified in ISO 226:2003 [39]. The human hearing frequency range relevant to the AR/MR use case considering 40-dB background noise. "Masked Zone" of gray area is approximation of equal-loudness at 40 dB with inverse A-weighted characteristics, and the boundary of "Pain Zone" and "Audible Zone" is approximated at 120 dB.

For example, background noise is roughly 40-dB SPL in a residential environment at night, roughly 50-dB SPL in a workplace, and roughly 60-dB SPL in shops [14]. The boundary between the audible zone and the masked zone is defined as the inverse characteristic of the A-weighted characteristic [40], assuming 40-dB of background noise.

In this paper, $P_{\mathrm{W}}(f_k)$ is assumed to be the sensory stimulus of the initial power considered, with the auditory scaling of the sound-masking effect and equal-loudness dynamic range. $P_{\mathrm{W}}(f_k)$ has a wider dynamic range at mid-range frequencies, and has a narrower range at both lower and higher frequencies, compared to ordinary $P(f_k)$.

4.4. Decay Time Scaling

Decay time is the reverberation time for each frequency, which is close to $RT60$. For example, if the reverberation time $RT60$ is 1 s, decay times are around 1 s. Therefore, assuming that the reference stimulus is $RT60$, and supposing that $K = 1$, $S_0 = RT60$ as follows:

$$T_{\mathrm{dr}}(f_k) = \ln \frac{T_{\mathrm{d}}(f_k)}{RT60}, \tag{4}$$

where $T_{\mathrm{dr}}(f_k)$ is the logarithm of the decay time ratio.

5. IMPLEMENTATION

In this section, we describe the implementation of each component in our system. As depicted in Fig. 1, the system synthesizes the reverberate IR from an image of a scene. Table 2 presents pseudo code for the synthesis of reverberation IR. Section 5.1 outlines how to look up the base waveforms from reverb templates by classifying the scene in an image. In Sect. 5.2, we explain the selection method of base waveforms to be adjusted. Section 5.3, we describe the application of CNNs to reverberate regression. Finally, Sect. 5.4 introduces the methods used to synthesize the IR.

5.1. Place Estimation

A CNN, which has been pre-trained on Places365, is used to estimate the scene in an image. The architecture of the CNN is Places365-ResNet [28]. In this processing block, the image is input and, depending on output scores from the CNN, labels of the estimated places are converted into categories from the Level 2 hierarchy. When the classified label is only "Outdoor-Natural man-made," the second candidate label is regarded as the estimated scene, for the reason described in Sect. 3.

Table 2 Pseudo code for synthesis of reverberation IR.

Input: $\boldsymbol{M} \Leftarrow$ Read-RGB-Image
Output: $\hat{\boldsymbol{h}}$ # adjusted IR

$i \Leftarrow$ PlaceEstimation($\boldsymbol{M}$)
$\boldsymbol{h} \Leftarrow$ ReverbTemplate(i)
$\hat{\boldsymbol{T}_{\mathrm{dr}}}, \hat{\boldsymbol{P}_{\mathrm{W}}}, \hat{RT60p} \Leftarrow$ ReverbReg($\boldsymbol{M}$)
$\hat{\boldsymbol{T}_{\mathrm{d}}}, \hat{\boldsymbol{P}} \Leftarrow$ ScaleConversion($\hat{\boldsymbol{T}_{\mathrm{dr}}}, \hat{\boldsymbol{P}_{\mathrm{W}}}, \hat{RT60p}$)

ReverbSynthesis: # see Appendix
$\boldsymbol{T}_{\mathrm{d}}, \boldsymbol{P} \Leftarrow$ CalculateEDR($\boldsymbol{h}$)
$\boldsymbol{\lambda}_{\mathrm{mod}} \Leftarrow$ CalculateDecayParameter($\boldsymbol{T}_{\mathrm{d}}, \hat{\boldsymbol{T}_{\mathrm{d}}}$)
$\boldsymbol{a}_{\mathrm{mod}} \Leftarrow$ CalculatePowerParameter($\boldsymbol{P}, \hat{\boldsymbol{P}}$)
$\hat{\boldsymbol{h}} \Leftarrow \boldsymbol{a}_{\mathrm{mod}} \times e^{-\boldsymbol{\lambda}_{\mathrm{mod}}} \times \boldsymbol{h}$

5.2. Sampling Waveforms

Since the echo density [41,42] affects the spatial impression, plausible waveforms for places have been selected. In this block, the IR, as representative reverberation waveforms, is looked up by inputting the place label. As a preliminary method in this paper, IRs have the median value of $RT60p$ for each scenes, which are regarded as the representative waveforms.

5.3. Reverb Regression

This stage performs regression of the reverberate parameters from the input image. Images in the database include those with compositions that are not suitable for estimating reverberation. For example, an image that only shows part of the floor or ceiling, or extremely zoomed-in images. After discarding such images, about 20,000 images were extracted. Of those, 2,826 images were used as test data; that is to say, these images were not used as training data.

Since some IRs are associated with multiple images taken from the same position but with different compositions, the number of IRs is not equal to the number of images. The output parameters for training were $RT60p$, $P_{\mathrm{W}}(f_k)$, and $T_{\mathrm{dr}}(f_k)$.

For normalization of the data in preprocessing, output parameters were subtracted from the mean of the differences and divided by the standard deviation of the differences. Since $P_{\mathrm{W}}(f_k)$ and $T_{\mathrm{dr}}(f_k)$ are frequency distributions, they were calculated for each frequency. To eliminate the influence of luminance, the parameters of the input images were subtracted from the mean and divided by the standard deviation for each RGB. The images were 224×224 pixels in size.

As [8] refers to applying CNNs [26] to regression, our regression architecture of CNNs is an extension of the ResNet 50-layer [27] network of Imagenet examples in Chainer. However, we added a fully connected final layer of 21 neurons. Training had the following settings: an epoch size of 75, a batch size of 75, and the loss function was the mean squared error. The model was run using the Chainer deep-learning framework on dual Nvidia GTX 1080s processors.

5.4. Reverb Synthesis and Processing

This phase edits the base IR output from the place estimation stage, according to the output parameters of the

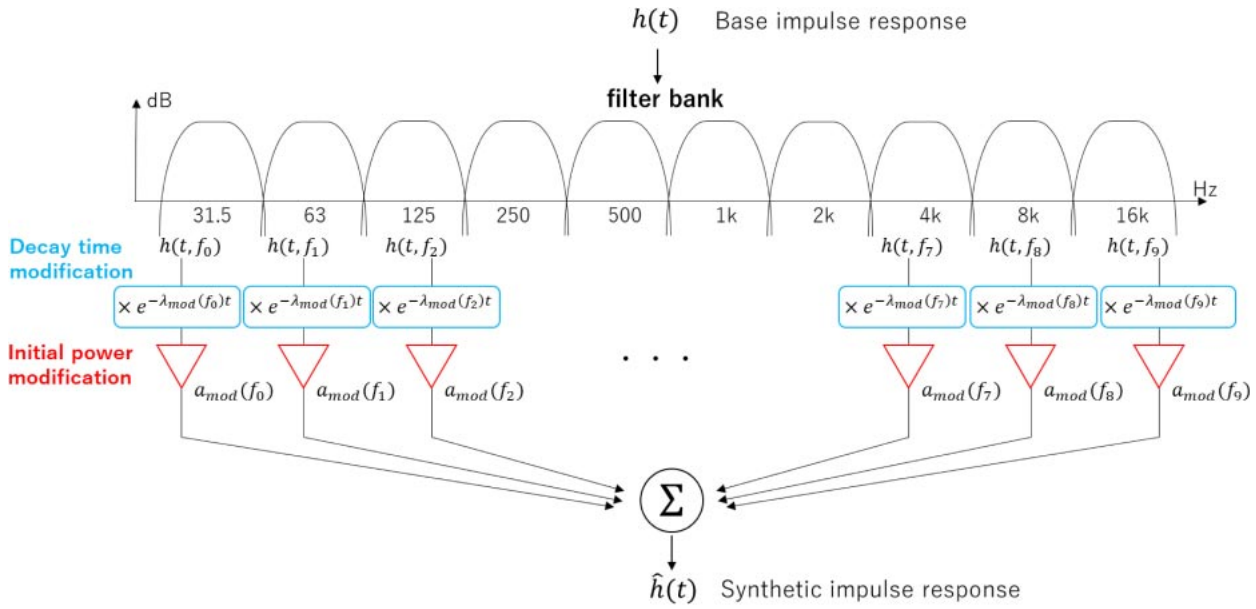


Fig. 4 Method for synthesizing a reverberation IR with frequency-dependent reverberation time and initial power [10,11,20].

reverberation characteristics estimation. The base IR was synthesized to include the characteristics of the reverb regression block outputs. Figure 4 depicts the method for adjusting the reverberation characteristics in each frequency band [10,11,20].

Supposing that $\hat{h}(t)$ is the reverberation IR that has been adjusted to the estimated characteristics, then

$$\hat{h}(t) = \sum_{k=0}^{9} a_{\mathrm{mod}}(f_k) \cdot e^{-\lambda_{\mathrm{mod}}(f_k)t} \cdot h(t, f_k), \qquad (5)$$

where $h(t, f_k)$ denotes the base IR divided by the frequency band f_k, $a_{\mathrm{mod}}(f_k)$ is a parameter for adjusting initial power, and $\lambda_{\mathrm{mod}}(f_k)$ is a parameter for adjusting decay time.

Originally, this method assumes that input is a stationary signal such as white noise. Calculating the adjustable parameters should consider that the input signal already has a decaying envelope characteristic since the input is reverberation IR (see Appendix).

Since the edited IR has the coefficients of the FIR filter, we implemented the 32-bit float FIR filter, which had a coefficient length of 480,000 (10 s with 48 kHz sampling rate). When the reverberation time is over 10 s, the length of the IR is truncated [21].

6. RESULT AND EVALUATION

In this section, we evaluate the accuracy of the regression. As a performance comparison, we compared our proposed auditory scaling method (referred to as Ours) with a regression of linear parameters that doesn't use auditory-characteristic scaling (referred to as Linear). The Linear approach is representative of conventional reverb synthesis methods [8]. For each method, we evaluated the accuracy of the estimation of reverberate parameters with the classified place. For each error comparison, a Mann-Whitney U test was also performed [43,44].

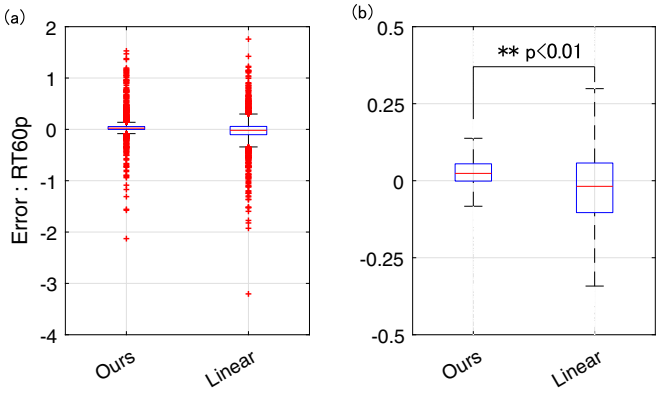


Fig. 5 Boxplots of the $RT60p$ regressions. "Ours" is proposed auditory scaling method, and "Linear" is the conventional method [8]. (a) displays the whole. dataset. (b) zooms in on the mean values and omits outliers.

Calculated reverberation times are evaluated in terms of the error in $RT60p$ from auditory scaling. Figure 5 depicts the box plots of this error.

Comparing Ours and Linear, reveals a significantly lower error in the proposed method ($p < 0.01$). Excluding outliers, the minimum and maximum values in the auditory scaling results are -0.06 and 0.08, and are 0.94 and 1.08 on the linear scale. The extremes in Linear are -0.34 and 0.29, which are linearized to 0.71 and 1.37. The error of $RT60p$, when converted to a linear scale indicates the ratio between the reference value and the estimated value.

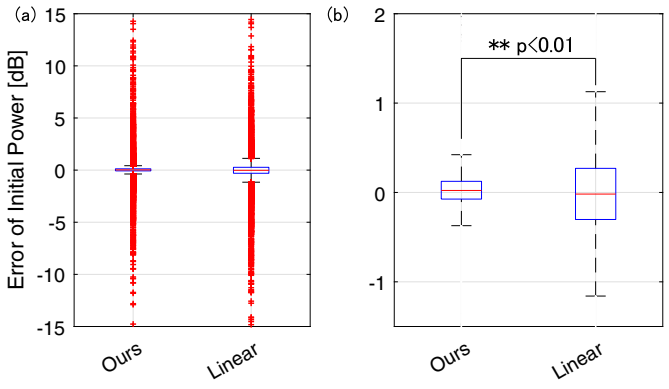


Fig. 6 Boxplots of the initial power regressions. "Ours" is proposed auditory scaling method, and "Linear" is the conventional method [8]. (a) displays all results. (b) zooms in on the central values.

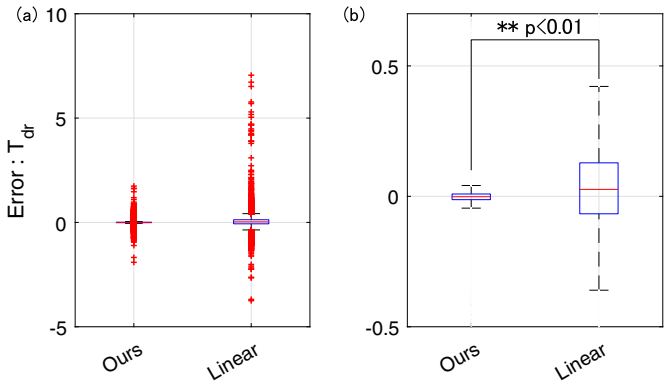


Fig. 7 Boxplots of the decay time regressions. "Ours" is proposed auditory scaling method, and "Linear" is the conventional [8]. method. (a) displays the all results. (b) zooms in on the central values.

Figures 6 and 7 present box plots of overall errors at all frequencies in initial powers and decay times. Calculated decay times are evaluated in terms of the error in T_{dr} from auditory scaling. Errors with the proposed method are significantly lower ($p < 0.01$), which corresponds with the result of reverberation time.

Therefore, compared to the conventional method, the proposed auditory scaling method appears to be effective for the regression of reverberation parameters.

7. USER EVALUATION

This study also evaluated whether the reverberation sound generated by our auditory scaling method is suitable for the impression of the scenery image, particularly in comparison with the conventional linear scaling method.

7.1. Experimental Condition

To evaluate perceptual impression, a framework within the MUltiple-Stimulus with Hidden Reference and Anchor (MUSHRA) method [45] was employed, as a method for quantifying subjective assessments. In the experiment, MUSHRA was intended to provide a reliable absolute measure of audio quality. No reference stimulus was used in this study; instead, whether each stimulus was plausible reverberation for the shown image was relatively evaluated. User evaluation was completed using the webMUSHRA [46] tool, the evaluation user interface of which is depicted in Fig. 8, along with evaluation images from webMUSHRA [46].

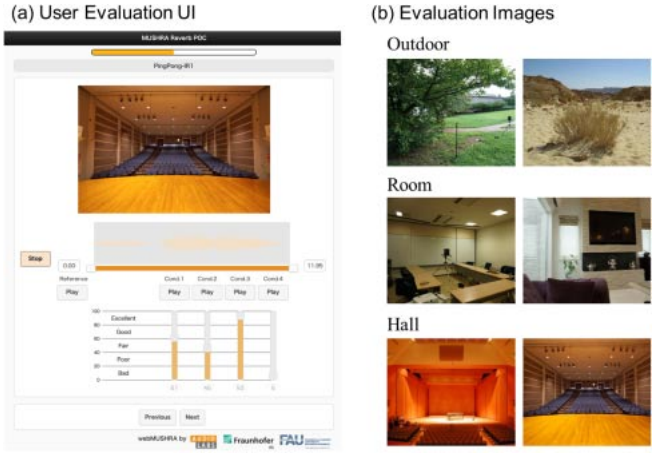


Fig. 8 (a) The WebMUSHRA [46] framework was used to present a user evaluation test. Four types of reverberation for each scene were evaluated. (b) displays evaluated images.

7.1.1. Stimuli

We compared our auditory scaling method (referred to as Ours), the measured IRs (referred to as Real), the conventional linear estimation method (referred to as Linear) and dry sounds (referred to as Dry). As such four types of reverberant sound (referred to as Method) were evaluated. The sound source was a ping-pong ball falling onto the floor, as this has an easily recognizable reverberation effect. The length of Dry sound source was about three seconds. Evaluation stimuli sources were obtained by convolution of the IR, which is synthesized or prepared with each method. These stimuli were randomly assigned to each playback button in each trial.

7.1.2. Scenes

We prepared three scenes (referred to as Scene), which were a large space (referred to as Hall), a narrow space (referred to as Room), the outdoors (referred to as Outdoor). The reverberation time was longer in the order of Hall, Room, and Outdoor. There were two images for each scene, meaning that a total of six images were evaluated. These images were displayed to subjects in a random order.

7.1.3. Subjects

Fourteen subjects between the ages of 22 and 42 participated in the experiment, all of whom exhibited

normal hearing ability, had no history of hearing problems, and were not audio-engineering professionals.

7.1.4. Procedures

Subjects were required to quantify the perceived impression on a numerical rating scale ranging from 0 to 100 with slider UI. A score of 100 indicated that the sound stimuli matched the image of the scene very well, whereas a score of 0 indicated an extremely unnatural match. Each set had one image and four types of evaluation sounds. After the instructions, but before the subject embarked on the actual evaluation, a test session was conducted. The evaluation of this test material was not included in the test results, but it confirmed that there was no problems with operation or understanding.

At the start of the assessment, subjects were given the following detailed instructions:

- Imagine you are in the space of the displayed image, and evaluate whether each reverberation matches that space.
- Listen to all test sound by clicking on the respective buttons above each slider.
- You may listen to the test materials in any order and any number of times.
- Your rating of the audio quality of each test sound should be indicated with the respective slider.
- Your rating should be higher scores for sounds that match the image, and lower scores for sounds that do not match the image.
- When you are satisfied with your ratings for all the test materials, click the "Next" button at the bottom of the screen.
- The evaluations are complete when you have finished evaluating six images.

7.1.5. Equipment and environment

Subjects listened to the stimuli with open-type headphones (Sennheiser HD660S), which were connected to a headphone amplifier (Sony PHA-2) and a PC (Apple Macbook Air) via USB. The experiment was conducted in a quiet room. The background noise level in the experimental environment was measured with a sound level meter (YCT YC-30), at roughly 36 dB to 45 dB.

7.2. Results and Discussion

We performed analysis of variance (ANOVA) [47,48] using Tukey's post-hoc test in order to obtain multiple comparisons for the assessment of differences between group means in the experimental results.

Figure 9 depicts the results of the subjective scores by method. The results of the ANOVA are shown in Table 3. There were significant differences in the mean values among all methods ($p < 0.01$).

Figure 10 depicts the results for each scene. In Outdoor, there was no significant difference between Dry and Linear ($p = 0.79$), but there were significant differences ($p < 0.01$) between every other combination of groups in this category. With regards to Room, there was no significant difference between Linear and Ours ($p = 0.39$), but there were significant differences ($p < 0.01$) between every other combination of groups in this category. The results for Hall reveal no significant difference between Ours and Real ($p = 0.32$), but significant differences ($p < 0.01$) between every other combination of groups in this category.

The score for Linear in Outdoor was lower than that of Room or Hall, and the score for Dry in Outdoor was higher than that of Room or Hall. Moreover, several subjects scored Dry higher than Linear in Outdoor, and there was a large variation between scores for Ours in Outdoor. These patterns suggests that more accurate estimation is not only possible, but also necessary for outdoor sounds, which have extremely short reverberation time characteristics. These tendencies are consistent with the notion that the smaller the stimulus, the greater the amount of change in perceived intensity, as the Weber-Fechner logarithm in Eq. (2) denotes.

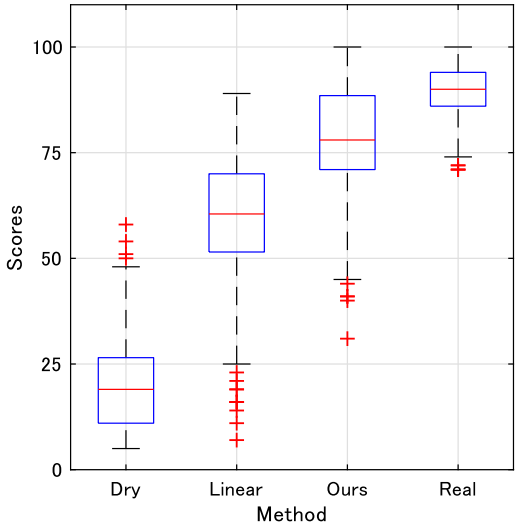


Fig. 9 Comparison of the subjective scores for different methods (Method) of reverberation; Dry is dry sound or without reverberation, Linear is conventional linear scale estimation, Ours is proposed auditory scaling method, Real is the actual measured IR. There were significant differences in the mean values among all methods ($p < 0.01$).

Table 3 ANOVA table of subjective score.

Source	Sum Sq.	d.f.	Mean Sq.	F
Method	257,501.1	3	85,833.7	890.58*
Method × Scene	30,746.4	2	5,124.4	53.17*

*: $p < 0.05$

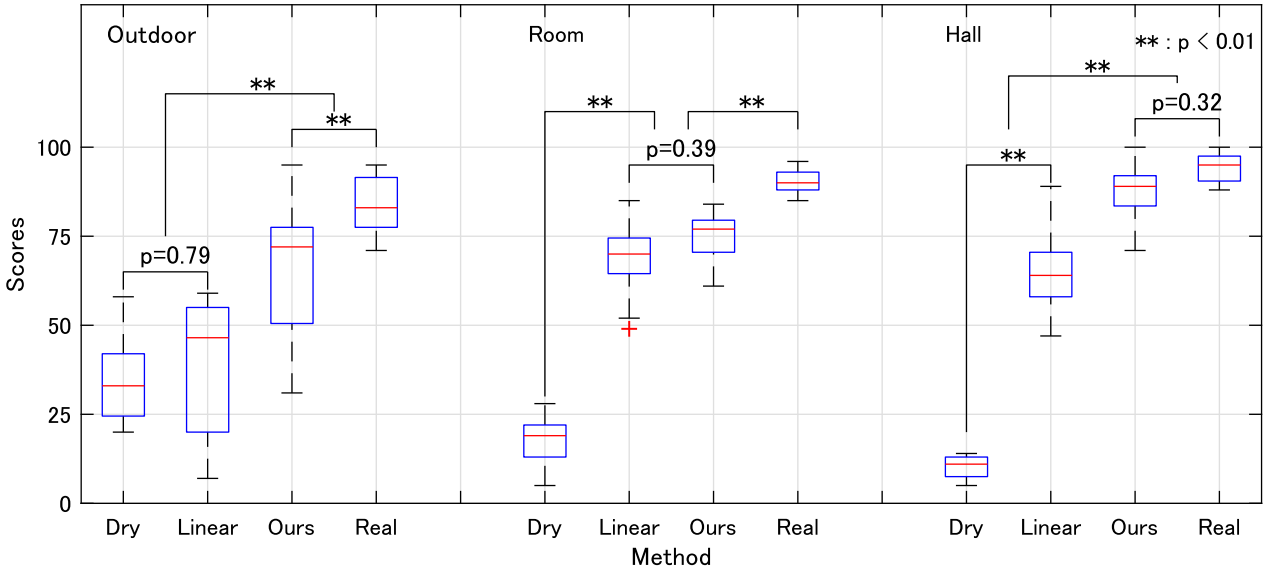


Fig. 10 Comparison of the subjective scores for different methods (Method) of reverberation by scene (Scene).

Comparing Ours and Linear reveals no significant difference between scores for Room, but the score of for Ours was generally higher than that for Linear. Furthermore, there was no significant difference between Ours and Real in Hall. These trends are also consistent with features of the Weber-Fechner law that is denoted in Eq. (2). This analysis reinforces the conclusion that Ours is effective for characteristics with long reverberation times.

8. CONCLUSIONS AND FUTURE WORK

This study proposes an auditory scaling method to reduce the estimation error of reverberation characteristics with regards to estimating reverberation characteristics from a 2D image. This method was evaluated through user tests and analysis, which have allowed us to draw the following conclusions: draw the following conclusions:

- Auditory scaling improves the estimation accuracy of reverberation characteristics.
- Subjective evaluation scores were generally higher for the auditory scaling method than for the conventional linear scale method.
- The results suggest that Ours is effective for characteristics with long reverberation times.

The observations reported in this study are useful for future advances to AR/MR applications. A wider range of applications may be developed using our method, which add reverberant sounds to images and video. For example, natural reverberation may be added to the dubbed sound of a movie, or reverberation characteristics estimation may be utilized for dereverberation in a speech-recognition system. Future research directions also include application of our method to various reverberation estimation systems in addition to AR/MR systems.

REFERENCES

[1] S. Orts-Escolano, C. Rhemann, S. Fanello, W. Chang, A. Kowdle, Y. Degtyarev, D. Kim, P. L. Davidson, S. Khamis, M. Dou, V. Tankovich, C. Loop, Q. Cai, P. A. Chou, S. Mennicken, J. Valentin, V. Pradeep, S. Wang, S. B. Kang, P. Kohli, Y. Lutchyn, C. Keskin and S. Izadi, "Holoportation: Virtual 3D teleportation in real-time," *Proc. 29th Annu. Symp. User Interface Software and Technology*, pp. 741–754 (2016).

[2] J. M. Jot and K. S. Lee, "Augmented reality headphone environment rendering," *Proc. Audio Eng. Soc. Conf.: 2016 AES Int. Conf. Audio for Virtual and Augmented Reality*, 6 pages (2016).

[3] M. Barron and A. H. Marshall, "Spatial impression due to early lateral reflections in concert halls: The derivation of a physical measure," *J. Sound Vib.*, **77**, 211–232 (1981).

[4] A. Rungta, C. Schissler, N. Rewkowski, R. Mehra and D. Manocha, "Diffraction kernels for interactive sound propagation in dynamic environments," *IEEE Trans. Vis. Comput. Graph.*, **24**, 1613–1622 (2018).

[5] L. Antani, A. Chandak, L. Savioja and D. Manocha, "Interactive sound propagation using compact acoustic transfer operators," *ACM Trans. Graph.*, **31**(1), pp. 1–12 (2012).

[6] R. Mehra, C. Malloy, M. Lin, A. Rungta, C. Schissler and D. Manocha, "Syncopation: Interactive synthesis-coupled sound propagation," *IEEE Trans. Vis. Comput. Graph.*, **22**, 1346–1355 (2016).

[7] C. Schissler, C. Loftin and D. Manocha, "Acoustic classification and optimization for multi-modal rendering of real-world scenes," *IEEE Trans. Vis. Comput. Graph.*, **24**, 1246–1259 (2018).

[8] H. Kon and H. Koike, "Deep neural networks for cross-modal estimations of acoustic reverberation characteristics from two-dimensional images," *Proc. 144th AES Convention*, 10 pages (2018).

[9] H. Kon and H. Koike, "Estimation of late reverberation characteristics from a single two-dimensional environmental image using convolutional neural networks," *J. Audio Eng. Soc.*, **67**, 540–548 (2019).

[10] J. A. Moorer, "About this reverberation business," *Comput. Music J.*, 13–28 (1979).
[11] J. M. Jot, L. Cerveau and O. Warusfel, "Analysis and synthesis of room reverberation based on a statistical time-frequency model," *Proc. 103rd AES Convention*, 42 pages (1997).
[12] Z. Meng, F. Zhao and M. He, "The just noticeable difference of noise length and reverberation perception," *Proc. 2006 Int. Symp. Communications and Information Technologies*, pp. 418–421 (2006).
[13] F. Martellotta, "The just noticeable difference of center time and clarity index in large reverberant spaces," *J. Acoust. Soc. Am.*, **128**, 654–663 (2010).
[14] A. P. G. Peterson, *Handbook of Noise Measurement*, 9th ed. (GenRad, Concord, 1980), Chap. 2, pp. 3–16.
[15] G. A. Miller, "Sensitivity to changes in the intensity of white noise and its relation to masking and loudness," *J. Acoust. Soc. Am.*, **19**, 609–619 (1947).
[16] R. E. Fancher, *Pioneers of Psychology*, 3rd ed. (W. W. Norton & Company, New York, 1996), Chap. 4, pp. 135–173.
[17] V. Välimäki, J. Parker, L. Savioja, J. O. Smith and J. Abel, "More than 50 years of artificial reverberation," *Proc. Audio Eng. Soc. Conf.: 60th Int. Conf.: DREAMS* (*Dereverberation and Reverberation of Audio, Music, and Speech*), 12 pages (2016).
[18] W. G. Gardner, *Reverberation Algorithms* (Springer US, Boston, 2002), Chap. 3, pp. 85–131.
[19] M. R. Schroeder, "New method of measuring reverberation time," *J. Acoust. Soc. Am.*, **37**, 409–412 (1965).
[20] J. M. Jot, "An analysis/synthesis approach to real-time artificial reverberation," *Proc. ICASSP 92: 1992 IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Vol. 2, pp. 221–224 (1992).
[21] U. Zölzer, *Digital Audio Effects* (John Wiley & Sons, Hoboken, 2004), Chap. 6, pp. 137–200.
[22] S. Kirkup, *The Boundary Element Method in Acoustics* (Integrated Sound Software, Heptonstall, 2007), Chap. 1, pp. 1–19.
[23] S. G. Parker, J. Bigler, A. Dietrich, H. Friedrich, J. Hoberock, D. Luebke, D. McAllister, M. McGuire, K. Morley, A. Robison and M. Stich, "Optix: A general purpose ray tracing engine," *ACM Trans. Graph.*, **29**(4), pp. 1–13 (2010).
[24] S. Bell, P. Upchurch, N. Snavely and K. Bala, "Material recognition in the wild with the materials in context database," *Proc. 2015 IEEE Conf. Computer Vision and Pattern Recognition*, pp. 3479–3487 (2015).
[25] A. Krizhevsky, I. Sutskever and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Proc. 25th Int. Conf. Neural Information Processing Systems*, Vol. 1, pp. 1097–1105 (2012).
[26] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv:1409.1556*, 14 pages (2014).
[27] K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition," *Proc. 2016 IEEE Conf. Computer Vision and Pattern Recognition*, pp. 770–778 (2016).
[28] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, **40**, 1452–1464 (2018).
[29] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," *Proc. 2010 IEEE Comput. Soc. Conf. Computer Vision and Pattern Recognition*, pp. 3485–3492 (2010).
[30] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke and A. Rabinovich, "Going deeper with convolution," *arXiv:1409.4842*, 12 pages (2014).
[31] S. Yan, M. Lin and Q. Chen, "Network in network," *arXiv:1312.4400*, 10 pages (2014).
[32] Waves. IR1 Convolution Reverb. https://www.waves.com/plugins/ir1-convolution-reverb/, (accessed 1 Sept. 2018).
[33] AudioEase. ALTIVERB. https://www.audioease.com/altiverb/, (accessed 1 Sept. 2018).
[34] boom. OUTDOOR IMPULSE RESPONSES. https://www.boomlibrary.com/sound-effects/outdoor-impulse-responses/, (accessed 1 Sept. 2018).
[35] S. Shelley and D. T. Murphy, "Openair: An interactive auralization web resource and database," *Proc. 129th AES Convention*, 9 pages (2010).
[36] International Organization for Standardization, ISO 3382-1:2009, Acoustics — Measurement of room acoustic parameters — Part 1: Performance spaces, 26 pages (2009).
[37] O. Sheynin, "Fechner as a statistician," *Br. J. Math. Stat. Psychol.*, **57**, 53–72 (2004).
[38] H. Fletcher and W. A. Munson, "Loudness, its definition, measurement and calculation," *J. Acoust. Soc. Am.*, **5**, 82–108 (1933).
[39] International Organization for Standardization, ISO 226:2003, Acoustics — Normal equal-loudness-level contours, 18 pages (2003).
[40] International Electrotechnical Commission, IEC 61672-2:2013, Electroacoustics — Sound level meters — Part 2: Pattern evaluation tests, 80 pages (2013).
[41] P. Huang, J. S. Abel, H. Terasawa and J. Berger, "Reverberation echo density psychoacoustics," *Proc. 125th AES Convention*, 10 pages (2008).
[42] S. J. Schlecht and E. A. P. Habets, "Feedback delay networks: Echo density and mixing time," *IEEE/ACM Trans. Audio Speech Lang. Process.*, **25**, 374–383 (2017).
[43] J. D. Gibbons and S. Chakraborti, *Nonparametric Statistical Inference* (Chapman & Hall, London, 2010), Chap. 6, pp. 227–274.
[44] M. Hollander, D. A. Wolfe and E. Chicken, *Nonparametric Statistical Methods*, 3rd ed. (Wiley, Hoboken, 2013), Chap. 4, pp. 115–150.
[45] ITU Telecommunication Standardization Union, ITU Recommendation BS.1534-1, Method for the subjective assessment of intermediate quality level of coding systems, 18 pages (2015).
[46] M. Schoeffler, S. Bartoschek, F. R. Stöter, M. Roess, S. Westphal, B. Edler and J. Herre, "webmushra — a comprehensive framework for web-based listening tests," *J. Open Res. Softw.*, **6**(1), 8 pages (2018).
[47] W. J. Conover, *Practical Nonparametric Statistics*, 3rd ed. (Wiley, Hoboken, 1999), Chap. 3, pp. 39–114.
[48] M. R. Harwell, E. N. Rubinstein, W. S. Hayes and C. C. Olds, "Summarizing monte carlo results in methodological research: The one- and two-factor fixed effects anova cases," *J. Educ. Stat.*, **17**, 315–339 (1992).

APPENDIX: CALCULATION METHOD FOR ADJUSTED IR

This section outlines the calculation of $a_{\mathrm{mod}}(f_k)$ and $\lambda_{\mathrm{mod}}(f_k)$, which are adjustment parameters for the estimated IR.

The reverberation IR before the adjustment is denoted as $h_{\mathrm{b}}(t)$. The frequency bandwidth employed in this paper is a whole-octave band ranging from 31.5 Hz to 16 kHz.

The number of frequency band is 10 and the initial power and decay time of $h(t)$ are denoted as

$$\boldsymbol{P} = [P(f_0), P(f_1), \ldots, P(f_9)]^{\mathrm{T}} \quad \text{(A·1)}$$

$$\boldsymbol{T_d} = [T_d(f_0), T_d(f_1), \ldots, T_d(f_9)]^{\mathrm{T}}. \quad \text{(A·2)}$$

Supposing $\hat{h}(t)$ is an adjusted IR, then initial power and decay time are denoted as $\hat{\boldsymbol{P}}$ and $\hat{\boldsymbol{T_d}}$.

Since $T_d(f_k)$ is the time used to attenuate 60 dB from $P(f_k)$, the time-frequency envelope [11] is denoted as

$$ENV_h(t, f_k) \simeq P(f_k) - \frac{60}{T_d(f_k)} t \quad [\mathrm{dB}], \quad \text{(A·3)}$$

where $ENV_h(t, f_k)$ is the time-frequency envelope function.

Initial Power

Supposing that $a_{\mathrm{mod}}(f_k)$ is the turning parameter of initial power as depicted in Fig. 4, then Eq. (A·3) returns the following expressions of the time-frequency envelope of $\hat{h}(t, f_k)$:

$$ENV_{\hat{h}}(t, f_k) = 10 \cdot \log_{10} |a_{\mathrm{mod}}(f_k)|^2 + P(f_k) - \frac{60}{T_d(f_k)} t. \quad \text{(A·4)}$$

Substituting 0 for t in Eq. (A·3) yields the following expression:

$$ENV_h(0, f_k) = P(f_k). \quad \text{(A·5)}$$

Therefore, the adjusted initial power characteristics are denoted as

$$\hat{P}(f_k) = 10 \cdot \log_{10} |a_{\mathrm{mod}}(f_k)|^2 + P(f_k). \quad \text{(A·6)}$$

This equation can be rearranged as

$$a_{\mathrm{mod}}(f_k) = 10^{(\hat{P}(f_k) - P(f_k))/20}. \quad \text{(A·7)}$$

Decay Time

If $e^{-\lambda(f_k)t}$ is the envelope function in linear scale [11,20], we obtain the following relational expression:

$$10 \cdot \log_{10} |e^{-\lambda(f_k) \cdot T_d(f_k)}|^2 = -60 \quad [\mathrm{dB}], \quad \text{(A·8)}$$

$$\lambda(f_k) = \frac{C}{T_d(f_k)}, \quad \text{(A·9)}$$

where $\lambda(f_k)$ is the damping coefficient of the mode and $C = 3 \ln 10$.

Supposing that $b(t, f_k)$ is the stationary white noise signal through the whole-octave bandpass filter, introduces the following expression:

$$h(t) \simeq \sum_{k=0}^{9} c_{f_k} \cdot b(t, f_k) \cdot e^{-\lambda(f_k)t}, \quad \text{(A·10)}$$

where c_{f_k} is the bandpass energy adjustment coefficient.

If $\lambda_{\mathrm{mod}}(f_k)$ is the turning parameter of decay time as depicted in Fig. 4, $\hat{h}(t)$ is denoted as

$$\hat{h}(t) \simeq \sum_{k=0}^{9} c_{f_k} \cdot b(t, f_k) \cdot e^{-\lambda(f_k)t} \cdot e^{-\lambda_{\mathrm{mod}}(f_k)t}. \quad \text{(A·11)}$$

Since $b(t, f_k)$ and c_{f_k} are not related to the decay time, we obtain the following expressions:

$$e^{-\hat{\lambda}(f_k)t} = e^{-\lambda(f_k)t} \cdot e^{-\lambda_{\mathrm{mod}}(f_k)t} \quad \text{(A·12)}$$

$$\lambda_{\mathrm{mod}}(f_k) = \hat{\lambda}(f_k) - \lambda(f_k). \quad \text{(A·13)}$$

These equations can be rearranged as

$$\lambda_{\mathrm{mod}}(f_k) = C \cdot \left(\frac{1}{\hat{T}_d(f_k)} - \frac{1}{T_d(f_k)} \right). \quad \text{(A·14)}$$