Meyer-Kahlen, Nils; Schlecht, Sebastian; Lokki, Tapio

Perceptual roughness of spatially assigned sparse noise for rendering reverberation

Published in:
Journal of the Acoustical Society of America

DOI:
10.1121/10.0007048

Published: 01/11/2021

Please cite the original version:

This material is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of the repository collections is not permitted, except that material may be duplicated by you for your research use or educational purposes in electronic or print form. You must obtain permission for any other use. Electronic or print copies may not be offered, whether for sale or otherwise to anyone who is not an authorised user.
Perceptual roughness of spatially assigned sparse noise for rendering reverberation

Nils Meyer-Kahlen, Sebastian J. Schlecht and Tapio Lokki

Citation: The Journal of the Acoustical Society of America 150, 3521 (2021); doi: 10.1121/10.0007048
View online: https://doi.org/10.1121/10.0007048
View Table of Contents: https://asa.scitation.org/toc/jas/150/5
Published by the Acoustical Society of America

ARTICLES YOU MAY BE INTERESTED IN

More than just methods: Data and insights guide vowel acoustics research
The Journal of the Acoustical Society of America 150, R9 (2021); https://doi.org/10.1121/10.0006962

Tracking time differences of arrivals of multiple sound sources in the presence of clutter and missed detections
The Journal of the Acoustical Society of America 150, 3399 (2021); https://doi.org/10.1121/10.0006780

Direct discrete complex image method for sound field evaluation above a non-locally reacting layer
The Journal of the Acoustical Society of America 150, 3509 (2021); https://doi.org/10.1121/10.0007065

Double-pass consistency for amplitude- and frequency-modulation detection in normal-hearing listeners
The Journal of the Acoustical Society of America 150, 3631 (2021); https://doi.org/10.1121/10.0006811

REVIEWS OF ACOUSTICAL PATENTS
The Journal of the Acoustical Society of America 150, 3211 (2021); https://doi.org/10.1121/10.0006963

Mandatory dichotic integration of second-formant information: Contralateral sine bleats have predictable effects on consonant place judgments
The Journal of the Acoustical Society of America 150, 3693 (2021); https://doi.org/10.1121/10.0007132
Perceptual roughness of spatially assigned sparse noise for rendering reverberation

Nils Meyer-Kahlen,a) Sebastian J. Schlecht,b) and Tapio Lokki[c)
Department of Signal Processing and Acoustics, Aalto University, P.O. Box 13100, Aalto, Finland

ABSTRACT:
Multichannel auralizations based on spatial room impulse responses often employ sample-wise assignment of an omnidirectional response to form loudspeaker responses. This leads to sparse impulse responses in each reproduction loudspeaker and the auralization of transient signals can sound rough. Based on this observation, we conducted a listening test to examine the general phenomenon of roughness due to spatial assignment. First, participants assessed the roughness of both Gaussian noise and velvet noise, assigned sample-wise to up to 36 loudspeakers by two algorithms. The first algorithm assigns channels merely by selecting random indices, while the second one constrains the time between two peaks on each channel. The results show that roughness already occurs when few channels are used and that the assignment algorithm influences it. In a second experiment, virtualizations of the test were used to examine the factors contributing to increased roughness. We systematically show the effect of spatial assignment on noise and conclude that besides time-differences, level-differences caused by head-shadowing are the principal cause for the perceived roughness. The results have significance in spatial room impulse response rendering and spatial reverberator design. © 2021 Acoustical Society of America. https://doi.org/10.1121/10.0007048

(Received 10 May 2021; revised 20 September 2021; accepted 12 October 2021; published online 10 November 2021)
[Editor: Jonas Braasch] Pages: 3521–3531

I. INTRODUCTION
Understanding the perception of spatially sparse noise is a critical step towards improving the reproduction of reverberation. The most important example is found in the rendering stage of broadband parametric methods for spatial room impulse responses (SRIR), in which an omnidirectional impulse response undergoes sample-wise assignment to loudspeaker responses by using short-time broadband directional estimates of the SRIR. Such sample-wise directional estimates could be obtained, for example, by broadband intensity vector processing or time difference of arrival estimation. Such sample-wise assignment constitutes the rendering stage of the spatial decomposition method (SDM) (Tervo et al., 2013), and of other SRIR processing algorithms operating on broadband directional estimates, see, for example, Zaunschirm et al. (2020) or Göllés and Zotter (2020). Rendering sound by convolving signals with such sequences yields reverberators that sound smooth for continuous input signals. However, they may cause a “grainy” or “rough” impression of the reverberation when convolved with transient signals. So far, the effect of roughness due to spatial assignment has been noticed (McCormack et al., 2020) and a compensation strategy has been proposed (Gari et al., 2021), but no detailed analysis of the causes is available.

While broadband parametric SRIR processing is our main motivation, the aim of this study is to understand the perception and origin of roughness in spatially distributed and sparse noise sequences in general. The presented experiments use exponentially decaying Gaussian noise (GN), selected to mimic the late response of an actual SRIR, and also decaying velvet noise (VN). Velvet noise is a ternary, sparse sequence with special properties that has been used to create computationally inexpensive single-channel reverberators (Holm-Rasmussen et al., 2013; Järveläinen and Karjalainen, 2007; Välimaki and Prawda, 2021) and decorrelators (Alary et al., 2017; Schlecht et al., 2018). Until now, no spatial reverberators using velvet noise have been created, and we hope that our results can aid in such a design as well.

Throughout the study, we will use the term “roughness” to describe the perceptual quality of non-smooth sparse noise. The main reason is that it has been used in earlier studies regarding the perception of sparse noise by Järveläinen and Karjalainen (2007), Välimaki et al. (2013), and Välimaki and Prawda (2021). In psychoacoustics, roughness is linked to amplitude modulation (Fastl and Zwicker, 2007). A psychoacoustic measure of roughness, asper, is defined by the sensation caused by a 60 dB sinusoid at 1 kHz. 100% amplitude modulated with a frequency of 70 Hz. However, Fastl and Zwicker (2007) also explain that modulation does not need to be periodic to evoke the sensation of roughness, as for example a narrow band noise signal can be described as rough. In fact, the spatial assignment tested here can be interpreted as a case of pseudo-random amplitude modulation. Furthermore, when asked for
comments after each of our experiments, the subjects did not report major issues with using the attribute to rate our stimuli.

In the following, we describe two experiments. Experiment I was conducted in a multichannel loudspeaker array in an anechoic room. In its first part, participants were asked to rate the roughness of Gaussian noise assigned to different numbers of loudspeakers. In the second part, we tested the roughness of Gaussian noise and velvet noise, following assignment to loudspeakers with two different algorithms. Already this first experiment showed the principal effect of increased roughness due to spatial assignment.

Experiment II was designed to investigate the causes of this roughness in more detail. For this purpose, a headphone version of the test was created by convolving the stimuli with binaural room impulse responses (BRIR), measured in the same anechoic multi-channel reproduction chamber used before. Additional stimuli were created by removing time and level differences between the loudspeakers from the BRIR and including an omnidirectional response measured in the chamber.

First of all, we explain the connection to SRIR processing and revisit sparse noise sequences in Sec. II. Then, we define the used spatial assignment algorithms in Sec. III. Sections IV and V describe and discuss the two experiments conducted. In Sec. VI, we discuss implications for reverberation synthesis and possible future work in modelling. Section VII concludes the article.

II. BACKGROUND

A. Motivation from SRIR processing

The initial motivation for studying sample-wise assigned noise came from SRIR processing, where in the SDM and its variants (Gari et al., 2021; Tervo et al., 2013; Zaunschirm et al., 2020), every sample of an omnidirectional room impulse response is assigned to multiple loudspeaker responses based on instantaneous directional estimates. The underlying modelling assumption is that the entire room response can be decomposed into a collection of individual sound events that could be referred to as image sources (Tervo et al., 2013). While in the early part of the response, such a modelling assumption is appropriate, it is heavily violated in the late part of the response. Notwithstanding, the method has been applied successfully in many studies, including the comparison of concert halls (Lokki et al., 2016) and stage acoustics (Gari et al., 2019), smaller music venues (Tervo et al., 2015b), sound studios (Tervo et al., 2014), movie theaters (Riionheimo and Lokki, 2021), and cars (Kaplanis et al., 2017). However, recently it was reported that when rendering very transient sounds, artefacts are audible (Gari et al., 2021; Gölles and Zotter, 2020), which could be described as “roughness” or “graininess” (McCormack et al., 2020). It appears to be the sample-wise assignment that causes an otherwise smooth sounding tail of the room impulse response to sound rough. Remarkably this effect is inaudible when rendering continuous sounds. An example can be found online.1

In this study, we investigate the basic phenomenon of roughness due to short-term spatial assignment in isolation, aiming to understand the reason for its emergence. Therefore, we use artificial signals like Gaussian noise, instead of actual room impulse responses, mimicking the critical late part of the response, and assigning it uniformly over the sphere. We assume that this is the case were the strongest roughness should be audible. Furthermore, we have noticed a strong connection between the sparse sequences that occur after such spatial assignment of RIR, and artificial reverberators based on sparse sequences like velvet noise. In the following, such sequences are used to examine specific aspects of roughness in sample-wise assigned noise sequences as well.

B. Sparse noise sequences

Gaussian noise is temporally dense and fully described by the distribution of its sample values, as shown in the first row of Fig. 1. Just as a large number of random physical processes, the late part of a room impulse response can be modelled as a decaying Gaussian noise sequence (Moorer, 1979), arising from the central limit theorem (Badeau, 2019).

Velvet noise introduced by Järveläinen and Karjalainen (2007), on the other hand, represents an example of a ternary sparse sequence. The only possible non-zero values are −1 and 1, as in digital sequences such as maximum length sequences that were commonly used for measurements (Xiang, 2008) and even for binaural reverberation (Xiang et al., 2019). While digital sequences are dense, the value of velvet noise is zero most of the time: in the sequence shown in the last row of Fig. 1 for example, 95.83% of all values are zero. The sign of the non-zero values is selected randomly, with equal probability for both outcomes. In a variant of velvet noise called crushed velvet noise (Werner, 2019), different probabilities are assigned to positive and negative sign, in which more unipolar sequences have high-pass behaviour.

For any sparse noise sequence including velvet noise, one can define a mean pulse density $N$, measured in peaks per second. Its inverse gives the average gap size between peaks, $T$. Countless algorithms exist for generating ternary sparse sequences with the same mean gap size, but using different peak placement algorithms, also leading to different roughness perception (Vilimaki et al., 2013). It is important to remember that, independent of the peak placement, all of these sequences have the same value distribution. Therefore, the value distribution is insufficient for characterizing these sequences. However, one can statistically examine the placement of gaps between the peaks. Figure 1 shows VN and TRN as two examples.

The simplest approach for creating a temporally sparse sequence is to place samples with values −1 or 1 at only some randomly selected indices. The result is called total random noise (TRN) (Järveläinen and Karjalainen, 2007). It can be implemented by drawing $M$ unique random indices or by a procedure introduced in Rubak and Johansen (1999),
which is based on realizations $r(n)$ of a continuous, uniformly distributed random variable $R \sim \mathcal{U}(0, 1)$,

$$h(n) = \left\lfloor \frac{T_d}{T_d - 1} \left( r(n) - \frac{1}{2} \right) \right\rfloor,$$

(1)

where $\lfloor \cdot \rfloor$ denotes rounding to the closest integer, and $T_d = \bar{T}_d$ is the mean gap size in samples. Välimäki et al. (2013) found that even at high densities of $\bar{N} \geq 8000$ pulses per second ($\bar{T} \leq 0.125$ ms), TRN sounds rougher than Gaussian noise.

Interestingly, other generation algorithms can produce much smoother sounding sequences while maintaining the exact same mean gap size. Amongst these, velvet noise (VN) sounds especially smooth (Välimäki et al., 2013). It is generated by applying random jitter to a set of uniformly spaced peaks, so that the indices on non-zero samples can be computed as

$$k(m) = mT_d + \lfloor r(m)T_d \rfloor,$$

(2)

where $\lfloor \cdot \rfloor$ denotes rounding to the next lower integer.

At the determined indices $k$, an impulse with random sign is placed,

$$h(n) = \left\{ \begin{array}{ll} 2\lfloor r(m) \rfloor - 1 & n = k(m), \\ 0 & \text{else,} \end{array} \right.$$  

(3)

where $L$ is the number of loudspeakers and $\mathcal{U}\{1, L\}$ denotes a uniform discrete random variable with outcomes between 1 and $L$. We call such assignment unconstrained. After unconstrained assignment, the sequences on every single channel $h_l(n)$ have unbounded gap size, very much like the rough sounding total random noise sequence presented above, see Fig. 2.

Second, as an alternative assignment strategy, we form blocks of $L$ peaks and assign each peak in every block to exactly one loudspeaker. Formally written, block assignments $s_l : \{1, \ldots, L\} \rightarrow \{1, \ldots, L\}$ are selected from the set of random permutations of $L$ elements $S_L$, such that
We call this constrained assignment. It is analogous to the velvet noise generation using the jittered sampling scheme described above, in the sense that it also leads to a bounded distribution. Here, the bound is on the maximal number of peaks that pass before the next peak is assigned to a specific speaker; it can never be larger than $2L$. As a consequence, if the gap lengths in the original sequence were bounded, the gaps in the assigned channels are bounded too. An example of unconstrained and constrained assignment is shown in Fig. 2. Clearly, the gap distribution of each channel depends on the assignment principle. Figure 2 demonstrates the increased sparseness of the assigned channels, which have a mean gap size of 

$$\tilde{T}_L = L\tilde{T}.$$  

Note that while Fig. 2 shows the assignment of velvet noise, sample-wise assignment of Gaussian noise is done in exactly the same way. In this case, the assigned sequences become sparse, although the original sequence is dense.

IV. EXPERIMENT I: ASSIGNMENT TO LOUDSPEAKERS IN THE ANECHOIC CHAMBER

A. Method

The aim of experiment I.a was to test the roughness of Gaussian noise when assigned to a varying number of loudspeakers. Experiment I.b tested the different assignment algorithms. For this, both Gaussian noise and velvet noise were assigned to a subset of 36 loudspeakers using constrained and unconstrained assignments. The experiments were designed as multiple stimulus comparison tests. Ratings were obtained on a continuous scale with the lower end point “very rough” and the upper end point “as smooth as the reference.” Besides one button for the reference sound, the GUI showed one slider for each of the tested sequences, so that they could be compared directly.

All tested sequences had an exponentially decaying envelope with $T_{60} = 2$ s. The decay was applied to mimic the application in reverberators and the length was selected, such that it would be sufficiently long to provide enough realisations of the random generation variables, revealing the properties of the sequences. All experiments were carried out at a sampling rate of 48 kHz. The playback system consisted of a maximum number of 36 Genelec 8341 loudspeakers in an anechoic environment, fulfilling ISO 3745 down to 50 Hz. The system was calibrated such that a maximum of $L_{AE} = 72$ dB was measured in the center of the array. The test was implemented using Pure Data.

Nine participants (2 female, 7 male) took part in this first test. They were researchers and students of the Acoustics Lab, between 26 and 36 years of age, with a mean age of 28.9 years (SD = 4.31). We were interested in testing experienced listeners, who would be able to focus on the specific aspect of roughness. There is no unique definition of listener experience, so it is difficult to control precisely. However, Zacharov (2018) discussed several conventions. We considered participants that regularly take part in

$$s(m) = \sigma_{m/(L)}(m \mod L), \quad \sigma_i \in S_{L}.$$  

(6)
listening test as experienced listeners, and participants who design listening tests themselves as experts. According to this definition, all subjects were experienced, and five subjects could even be considered as experts. Seven participants had taken part in a velvet noise test before. One participant reported a reduced threshold of hearing at 4 kHZ, all others had no hearing loss known to their own knowledge. The authors did not take part in the test.

The experiment consisted of two tests: in experiment I.a, participants compared the roughness of a single-channel GN reproduced over one speaker to GN that was distributed to different numbers of loudspeakers. The seven stimuli and the reference were presented on one page, so that they could be compared. The loudspeaker positions are shown in Fig. 3. The hypothesis was that an increase in the number of channels leads to an increase in roughness.

In experiment I.b, participants were asked to rate GN and VN at two densities, following both constrained and unconstrained assignment to the full set of 36 loudspeakers. The seven stimuli and VN at two densities, following both constrained and unconstrained assignment to the full set of 36 loudspeakers. The seven stimuli and the reference were presented on one page, so that they could be compared. The loudspeaker positions are shown in Fig. 3. The hypothesis was that an increase in the number of channels leads to an increase in roughness.

In experiment I.b, participants were asked to rate GN and VN at two densities, following both constrained and unconstrained assignment to the full set of 36 loudspeakers. Again, the seven stimuli could be compared to each other. Here, the hypothesis was that the assignment algorithm accounts for a roughness difference. Additionally, we were interested in the overall roughness of the selected VN sequences in comparison to GN. Independently generated VN with \( T = 0.05 \) ms on all channels was also included. In pilot tests, participants found experiment I.b easier than I.a, as the spatial properties do not change. For this reason, the order of the tests was not randomized, but experiment I.b was always presented first. We would expect a larger variance in experiment I.a, if this test had been presented first.

### B. Results

For experiment I.a, Fig. 4(a) shows the result for Gaussian noise assigned to a varying number of loudspeakers. To confirm that interpretable differences are present in the data of experiment I.a, a one-way repeated measures ANOVA was conducted. Due to violation of the sphericity assumption according to a Mauchly test, \( \chi^2(20) = 32.29, p < 0.001 \). Greenhouse-Geisser corrected results are reported (\( \epsilon = 0.72 \)) (Howell, 2010). There was a significant effect of the number of loudspeakers; \( F(4.35, 112.99) = 388.96, p < 0.001 \). As post hoc tests, two-sided paired sample t-tests were conducted. The reported p-values were adjusted using the Bonferroni-Holm correction procedure (Holm, 1979).

In both experiments I.a and I.b, a single-channel GN sequence was used as a reference, and the unconstrained VN with \( T = 0.05 \) ms was included, such that the results of both parts can be related. The data shows the increase in roughness with an increasing number of loudspeakers. There was a significant effect between the two loudspeaker \( M = 91.26, SD = 11.80 \) and the four loudspeaker condition \( M = 70.08, SD = 17.80 \), \( t(26) = 5.90, p < 0.001 \). Also, the roughness increase that was perceived when assigning the noise to eight \( M = 49.59, SD = 17.04 \) instead of four loudspeakers \( M = 70.08, SD = 17.80 \) was significant, \( t(26) = 3.86, p = 0.002 \). Conversely, the mean difference between eight and 14 loudspeakers was only 8.53, and between eight and the maximally tested number of 36 loudspeakers, none of the small roughness differences were found to be significant.

The results of experiment I.b are shown in Fig. 4(b). Note that unconstrained GN on all 36 loudspeakers appeared in both tests. Again, after detecting a violation of sphericity using the Mauchly test, \( \chi^2(20) = 45.65, p < 0.001 \), the repeated measures ANOVA was conducted with Greenhouse-Geisser correction (\( \epsilon = 0.64 \)). It indicated significant differences between the conditions; \( F(3.81, 99.07) = 388.96, p < 0.001 \). As in experiment I.a, the distributed sequences are not perceived as smooth. The results for the VN sequence are very clear: although, for single-channel VN, a mean gap size of 0.5 ms is sufficient for a smooth sound when compared to GN (Järveläinen and Karjalainen, 2007; Välimäki et al., 2013), it no longer is after spatial assignment.

Moreover, there are significant differences between the algorithms used for spatial assignment. For the dense VN
sequence the difference between constrained assignment $(M = 93.40, SD = 8.63)$ and unconstrained assignment $(M = 64.76, SD = 15.89)$ is the largest, $t(26) = 8.23, p < 0.001$. Also, in the case of GN, there is a notable difference between constrained $(M = 48.91, SD = 13.49)$ and unconstrained assignment $(M = 36.27, SD = 10.46)$, $t(26) = 3.55, p = 0.003$. Even for the very sparse VN sequence, an effect was found when comparing constrained $(M = 6.35, SD = 6.96)$ and unconstrained $(M = 2.31, SD = 4.95)$ assignment, $t(26) = 3.28, p = 0.003$.

### C. Discussion

The results of experiment I contradict the inherent assumption made by broadband parametric SRIR rendering (Tervo et al., 2013). It assumes that since the individual channel perfectly sum up to the original sequence,

$$h(t) = \sum_l h_l(t), \quad (8)$$

no roughness should be heard if the original sequence sounds smooth. According to this, when the original sequence is Gaussian or sufficiently dense velvet noise $(T \leq 0.5\text{ ms})$, no roughness should be perceived at all. In experiment Ia, we see that the GN sequence does not sound smooth at all when assigned to multiple loudspeakers. Also, in experiment Ib, we see that velvet noise with $T \leq 0.5\text{ ms}$ sounds very rough after assignment. Further, the assignment algorithm does lead to significant differences, which should not be the case if the perfect sum was relevant for perception.

As an alternative explanation, one could assume that listeners would be able to use their binaural hearing system in order to focus on each channel individually, in a process of spatial release from masking (Litovsky, 2012). In that case, roughness perception would only depend on the signal of every assigned channel seen in Fig. 2. However, according to this principle, only the independent velvet noise should sound smooth in experiment Ib, as every channel has a mean gap size of $T_L = 0.5\text{ ms}$—the limit for a smooth-sounding velvet noise sequence found by Välimäki et al. (2013). Yet constrained VN with a mean gap size of $T = 0.041\text{ ms}$ before assignment is in fact given very high ratings, regardless of the fact that every channel has a density of only $T_L = LT = 36 \cdot 0.0416 = 1.5\text{ ms}$; thus it is only a third as dense as required for each one of them to sound smooth on their own.

Furthermore, it can be observed that after spatial assignment, the GN sounds rougher than the spatially assigned VN sequence with a mean gap size of $T = 0.041\text{ ms}$, although every channel of the GN is still more than 5 times denser than the VN (assigned Gaussian noise has a mean gap length of $T = 0.75\mu\text{s}$ on each channel). However, as it follows the Gaussian value distribution, many sample values are small. This shows that both values and gaps play a role in roughness perception.

As neither the summed sequence nor the individual sequences can account for the results, these observations demand a new explanation for the cause of roughness due to spatial assignment. Clearly, the final percept depends on the summed signal at the listeners ears, after assignment and propagation. Two influences on the propagation path are easily imaginable. First, even for a central listening position, there are small time differences $\tau_{i(L,R)}$ between the signals reaching the left and right ear of the listener $b_{i(L,R)}$. Second, head-shadowing causes direction dependent level modifications $g_{i(L,R)}^{(L,R)}$ in each ear signal

$$b_{i(L,R)}(t) = \sum_l g_{i(L,R)}^{(L,R)} h_l(t - \tau_{i(L,R)}). \quad (9)$$

This would mean that distributed velvet noise at the listener’s ears does not correspond to either of the sequences shown in Fig. 2, but would have characteristics similar to Fig. 5. These distributions have been calculated by assuming a distance of 9 cm from the center in the selected
A. Method

For this experiment, the loudspeaker setup was binauralized by convolving the loudspeaker signals with BRIRs, measured in the loudspeaker array using a KEMAR head and torso simulator (HATS). A diffuse-field equalizer derived from the mean of the magnitude responses from all directions was applied. Experiment II.a used the same signals as experiment I.b, in order to check the agreement between the loudspeaker and the headphone version of the test.

Thanks to the virtualization of the test, in experiment II.b, it was also possible to experiment with modified signals that would be impossible in a loudspeaker test. For this, a measurement with a GRAS 46AF omnidirectional measurement microphone was conducted. The microphone was placed roughly 9 cm away from the center of the array. The omnidirectional RIR allowed us to study the effect of small time differences on the roughness of the spatially distributed noise without the shadowing effect of the human head. An additional condition was created by time-aligning these responses, for which sample-valued time lags between the loudspeakers were computed using the cross correlation method. The measured and the time-aligned omnidirectional responses are shown in Fig. 6.

Also, the BRIR measurements were modified to create additional conditions. First, a diotic condition was created by mapping the responses measured at the left ear to both ears. Second, time alignment was applied to this diotic response. Last, for one condition, the levels of the loudspeakers were adjusted in the diotic response, based on the RMS within the first 10.5 ms of the response. As a test signal, unconstrained VN with a mean gap size of $T = 0.25$ ms was used. All in all, this resulted in six conditions, which are omnidirectional, omnidirectional with time-alignment, binaural, diotic, diotic with time alignment, diotic with level alignment. Also, binauralization of a sparse distributed VN sequence with $T = 0.5$ ms was included, so that the test would be anchored to the same high-roughness condition as before. During the pilot phase, it became clear that when applying the modifications, some of the sequences sounded smoother than the GN reference. In order to still resolve differences on the upper end of the scale, the GN reference was replaced with the perfect sum of all channels. Participants were instructed to give the highest rating, if they thought that any sequence sounded smoother than the reference. Every test was presented twice in random order.

The test was conducted using headphones, and participants were able to complete it in their home office. In total, 21 (19 male, 2 female) participants took part in the test. The mean age of participants was 29.62 years (SD = 3.15). All, except two, participants were experienced listeners, according to our definition above, 15 of them can be considered experts. 13 participants indicated having taken part in a VN test before. One participant indicated a reduced threshold of hearing at 2 kHz. In the questionnaire preceding the test, participants were also asked to indicate the headphone model they used for the test. All participants used high quality over-ear headphones produced by different manufacturers. The test was implemented using MATLAB. In this way, the fact that the test had to be conducted remotely introduces a large number of uncontrollable variables.

V. EXPERIMENT II: BINAURAL LISTENING TEST USING HEADPHONES

During the pilot phase of the first experiment, we have made an interesting observation: when placing an omnidirectional microphone in the center of the array and listening to its output outside of the reproduction chamber, the sequences sounded almost perfectly smooth. When placing a dummy head in the array and listening to the signal, they still sounded rough. This observation motivated experiment II, which aims at assessing the cause of the roughness in more detail.
B. Results

Before conducting a repeated measures ANOVA on the data from experiment II.a, a Mauchly test was run. It showed that the sphericity assumption is violated in this case as well, $\chi^2(20) = 113.4, p < 0.001$. The subsequent repeated measures ANOVA using Greenhouse-Geisser correction ($\epsilon = 0.57$) indicated significant differences, $F(3.42, 140.42) = 265.65, p < 0.001$.

For Gaussian noise, the difference between constrained ($M = 58.79, SD = 20.16$) and unconstrained ($M = 46.67, SD = 20.68$) assignment is significant again, $t(41) = 8.23, p < 0.001$. The same holds true for the denser version of constrained velvet noise ($M = 89.74, SD = 17.02$) and unconstrained velvet noise ($M = 74.5, SD = 21.93$), $t(41) = 3.85, p < 0.001$. However, also a general trend towards less roughness is observed and the variance is larger. Moreover, the difference between the sparser velvet noise assignment variants does not show a significant difference. Since a Kolmogorov-Smirnov test indicated that the difference of these two border cases does not follow a normal distribution, $D(41) = 0.25, p = 0.009$, a paired t-test is not a good choice. A Wilcoxon signed rank test also indicated no significant effect, $Z = -0.008, p = 0.92$. Although it seems that the headphone evaluation method makes it possible to compare the most apparent differences, the increased number of uncontrollable variables, as well as non individualized HRTFs make it more difficult to find effects.

Nevertheless, new insights are provided by the results obtained for the modified responses in experiment II.b, shown in Fig. 7(b). The repeated measures ANOVA using Greenhouse-Geisser correction $\epsilon = 0.55$, following a positive Mauchly test, $\chi^2(27) = 121.12 p < 0.001$, indicates significant differences, $F(3.86, 158.26) = 326.11, p < 0.001$.

As expected from our informal observation, the sequences convolved with the omnidirectional response ($M = 76.93, SD = 20.37$) sounds much smoother than the one created using the binaural response ($M = 13.91, SD = 14.67$), $t(41) = 18.71, p < 0.001$. Moreover, time-aligning the omnidirectional responses ($M = 83.10, SD = 15.96$) yielded a small but significant reduction of smoothness, when compared to the omnidirectional response with no alignment ($M = 76.93, SD = 20.37$), $t(41) = 3.34, p = 0.005$. A similarly small effect is observed between the time-aligned diotic response ($M = 22.78, SD = 16.80$) and the normal binaural response ($M = 18.21, SD = 15.59$), $t(41) = 2.89, p = 0.012$. However, performing level alignment on the diotic response ($M = 45.60, SD = 20.74$) yields a much larger reduction of roughness, when compared to the diotic response with no alignment ($M = 22.78, SD = 16.80$), $t(41) = 9.90, p < 0.001$.

**FIG. 6.** (Color online) Direct sound from the impulse responses used for the different conditions in experiment II. In the online version, each color represents the response of one loudspeaker channel.
There is no significant difference between the diotic and the binaural response; $t(41) = 1.86, p = 0.69$. Also, it should be noted that no other sequence is capable of recreating the full smoothness of the perfectly summed velvet noise sequence.

C. Discussion

The results of experiment II.b show that roughness due to spatial assignment is mainly caused by the effect of level differences, as a result of head shadowing. Both versions using the omnidirectional RIR and the level aligned diotic condition result in significantly higher smoothness ratings than the diotic condition that includes the level differences. Intuitively, if some of the sparse sequences on the individual channels are affected more by head-shadowing than others, they “stand out” in the summed response as strong peaks, as some peaks do in Fig. 5. The time differences caused by the non-central omnidirectional receiver were shown to have a lesser effect than the level differences.

However, also the version using the omnidirectional microphone measurement sounds less smooth than the perfectly summed sequence. This could be due to non-ideal loudspeaker calibration and the increased microphone directivity at high frequencies. Another explanation is that the room, even being anechoic, influences the sonic character of the noise. Only the dry, perfect summation develops the extraordinary smoothness of VN, which is reported to have a different sonic quality than Gaussian noise. In case of the binaural stimulus, despite the increase in smoothness through level alignment, the sequence could not lead to perfect smoothness. This is likely due to the simple nature of the broadband gain adjustment, and applying the inverse of each participants individual HRTFs magnitude response corresponding to the channels would probably decrease roughness even more.

Interestingly, the difference between the binaural and the diotic condition is relatively small. Thus, static binaural hearing mechanisms do not seem to play a large role in roughness perception of this kind. However, one participant indicated that comparing binaural and monaural stimuli was difficult. Also, dynamic cues might play a role, as the virtualized setting yielded less roughness than the loudspeaker condition in general. In experiment I, even though participants were instructed to stay seated, they could move their head and rotate their body. Head and body movements cause time-varying time and level differences and may increase roughness in this way.

VI. IMPLICATIONS AND FUTURE WORK

The results have implications for broadband SRIR methods, as well as for artificial reverberation design. With regards to SRIR methods, we have shown that roughness is an inherent property of sample-wise assignment that already occurs when assignment is performed to few loudspeakers. The effects caused by the listener’s head in the center of a loudspeaker array is sufficient to cause roughness, where level differences have a larger effect than time differences. This means that for mitigation, reproduction will have to go beyond assignment, making the individual responses less sparse, as it was done using allpasses by Gari et al. (2021) and though widening by Göllés and Zotter (2020). An interesting alternative would be to modify the value distribution of the assigned room impulse response, as we have seen that velvet noise, which does not have random value fluctuations can sound more smooth after assignment. In our experiment, assignment of the late part of a SRIR was mimicked by the Gaussian noise sequence, which has a white spectrum.

Note that if the original sequence is colored before spatial assignment, the spectrum is flattened by the process. In broadband parametric SRIR processing, this so-called
whitening effect needs to be compensated for by additional spectral correction (Tervo et al., 2015a; Zaunschirm et al., 2020), which is not discussed here. Also, it should be mentioned that the quality of parametric SRIR methods does not only depend on roughness, but also timbral and spatial accuracy need to be taken into account when evaluating such methods. Further, broadband methods are not the only possible choice when processing RIR, but methods with more elaborate sound field models that do not rely on sample-wise assignment have been described (McCormack et al., 2020).

The second field of application lies in artificial reverberation design with velvet noise. Recently, a new, efficient single-channel velvet noise reverberation algorithm was introduced (Välimäki and Prawda, 2021). We hope the insights into roughness caused by different assignment strategies can aid future developments into spatial velvet noise reverberation.

An excellent possibility for future work would be to predict roughness of sparse noise using a psychoacoustical model, for example, based on the work of Daniel and Weber (1997). It should be able to predict the required increase in a sequence when assigned to L loudspeakers, but ideally it should also allow for explaining the differing roughness perception between sparse noise generation algorithms, as observed by Välimäki et al. (2013).

VII. CONCLUSION

The presented experiments studied the roughness of spatially sparse noise. For the first experiment, we have used Gaussian noise as an artificial SRIR. The test has shown that perceived roughness increases when randomly assigning samples of Gaussian noise to multiple loudspeakers. The same is observed in the case of velvet noise. Furthermore, we have demonstrated that the introduced constrained assignment algorithm results in less roughness. With the listening experiments, we found that roughness is perceived already when assigning Gaussian noise to two loudspeakers, and increases until eight loudspeakers are reached. Constraining the assignment, such that the number of peaks that pass before the next peak is assigned to the same channel, leads to reduced roughness.

Thanks to a binauralized version of the test, we were also able to show that an important contributor to roughness due to spatial assignment is head-shadowing, which introduces level differences between the loudspeaker channels. The effect of level differences appears to be stronger than that of time differences. The role of binaural processing in roughness perception is more difficult to assess, but appears to be small, at least in static binaural reproduction.

ACKNOWLEDGMENTS

This research has received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 812719.

1Please find an example of a processed SRIR and all used stimuli at http://research.spa.aalto.fi/publications/papers/spatially_sparse_noise/


https://doi.org/10.1121/10.0007048


