S. Riedel, M. Frank, and F. Zotter, “The Effect of Temporal and
Directional Density on Listener Envelopment”
J. Audio Eng. Soc., vol. 71, no. 7/8, pp. 455–467, (2023 July/August).
DOI: https://doi.org/10.17743/jaes.2022.0088

The Effect of Temporal and Directional Density on Listener Envelopment

STEFAN RIEDEL,[1,*] *AES Student Member*, **MATTHIAS FRANK,**[1] *AES Associate Member* **AND**
(riedel@iem.at) (frank@iem.at)

FRANZ ZOTTER,[1] *AES Member*
(zotter@iem.at)

[1] *Institute of Electronic Music and Acoustics, University of Music and Performing Arts, Graz, Austria*

Listener envelopment refers to the sensation of being surrounded by sound, either by multiple direct sound events or by a diffuse reverberant sound field. More recently, a specific attribute for the sensation of being covered by sound from elevated directions has been proposed by Sazdov et al. and was termed listener engulfment. The first experiment presented here investigates how the temporal and directional density of sound events affects listener envelopment. The second experiment studies how elevated loudspeaker layers affect envelopment versus engulfment. A spatial granular synthesis technique is used to precisely control the temporal and directional density of sound events. Experimental results indicate that a directionally uniform distribution of sound events at time intervals $\Delta t < 20$ ms is required to elicit a sensation of diffuse envelopment, whereas longer time intervals lead to localized auditory events. It shows that elevated loudspeaker layers do not increase envelopment but contribute specifically to listener engulfment. Low-pass-filtered stimuli enhance envelopment in directionally sparse conditions, but impede control over engulfment due to a reduction of height localization cues. The results can be exploited in the technical design and creative application of spatial sound synthesis and reverberation algorithms.

0 INTRODUCTION

Listener envelopment (LEV) is a perceptual attribute to characterize the spatial impression of a sound field. It has been investigated by researchers in concert hall acoustics, spatial sound reproduction, and electroacoustic music [1–4]. According to Berg [5], various definitions have been used for envelopment, because the sensation is evoked either by room reverberation or surrounding direct sound events. The unifying factor seems to be the “sensation of being surrounded by sound” [5]. This generic definition has been adopted by various researchers in their works [3, 4, 6]. In contrast to apparent source width, which refers to the horizontal extent of an auditory event, LEV is related to the immersive auditory quality of a scene. It has been shown that LEV strongly correlates with the overall quality of the listening experience [7].

Previous work in the field of concert hall acoustics focused on the effect of early/late reverberation and its directional distribution. It suggested that late lateral energy is crucial for listener envelopment [8, 2, 9]. Studies additionally report correlation with reverberation from front, rear, and overhead directions [10–12] and the orchestral dynamics [13]. Literature on multichannel sound reproduction studied the required number of loudspeakers and their arrangement to optimally reproduce the spatial impression of a diffuse sound field. It was concluded that as few as four loudspeakers are sufficient in the case of low-pass noise signals or music stimuli, whereas more loudspeaker directions are required for broadband pink noise signals [14–16].

Most of the aforementioned studies varied the number of active loudspeakers and their directional distribution, and the sound stimuli were typically uncorrelated stationary noise signals or reverberated music signals. The required temporal density of sound events to elicit a sensation of diffuse envelopment remains uninvestigated. Literature on the processing lag of the binaural hearing mechanism reports time constants between 50 and 200 ms [17–20], suggesting that surrounding sound events at significantly shorter time intervals lead to a diffuse and potentially enveloping perception.

This motivates the following research questions: What is the required temporal density of surrounding sound events

[*] Partial results were presented at the 48th DAGA, Stuttgart, Germany, March 2022 [1]. To whom correspondence should be addressed, e-mail: riedel@iem.at. Last updated: March 20, 2023

to elicit a sensation of envelopment? Is the highest degree of envelopment elicited by a stationary and isotropic diffuse sound field, or by a sound field that exhibits audible spatio-temporal fluctuations/modulations? These questions are relevant in the technical design and creative application of artificial reverberators [21–25], spatial sound synthesis techniques [4, 26, 27], and spatial up-mixing algorithms [28, 29].

Only few experimental studies have been conducted on the spatial impression of 3D versus 2D sound fields, e.g., reporting on perceptual attributes such as "subjective diffuseness" [15, 30], "3D envelopment" [31], or overall listening experience [7]. Loudspeaker setups with height layers increased perceived diffuseness over "ear-height only" arrangements in an experimental study using pink noise signals [15]. A successive study on perceived diffuseness investigated the effect of the listener's head movements, and showed that "with-height" reproduction could only enhance auditory diffuseness if listeners were explicitly allowed/asked to tilt their head sideways, effectively moving height loudspeakers into the interaural axis [30]. To better describe the perceptual effect achieved by height loudspeaker layers, the term listener engulfment (LEG, "being covered by sound") was proposed by Sazdov et al. [32]. However, experimental data comparing the perception of envelopment and engulfment is limited [32, 4].

It seems that further experiments are necessary to clarify the perceptual effects of height layers. Therefore, this study additionally covers the following research questions: Do height layers enhance listener envelopment, or rather contribute to a distinct sensation (engulfment)? How are these attributes affected by the stimulus bandwidth (high-frequency content)?

Methodologically, to control the temporal and directional density of sound events in experimental conditions, a spatial granular synthesis technique is employed in this study. Recently, a similar technique used sample-wise assignment of noise signals in order to study perceptual roughness in spatial impulse response rendering and up-mixing [33]. Linear time-invariant (LTI), infinite impulse response reverberation algorithms such as feedback delay networks render an increasingly dense diffuse reverberation tail, which impedes the synthesis of stimuli with well-controlled temporal density of sound events. Nevertheless, recent developments enable directional control [24]. LTI finite impulse response (FIR) filters for spatial reverberation [25, 34] can be designed to control the spatio-temporal density of a sound field, but the output density will depend on the temporal character of the input signal (transient vs. stationary). An algorithm that directly assigns time-windowed audio signals to specified directions can synthesize both sparse and diffuse sound fields, given a temporally dense or stationary input signal. Therefore, a spatial granular synthesis is the preferred method in this study.

The first section of the paper describes and evaluates the method of spatial granular synthesis. The second section of the paper describes and discusses listening experiments, which used spatial granular synthesis to investigate the proposed research questions.

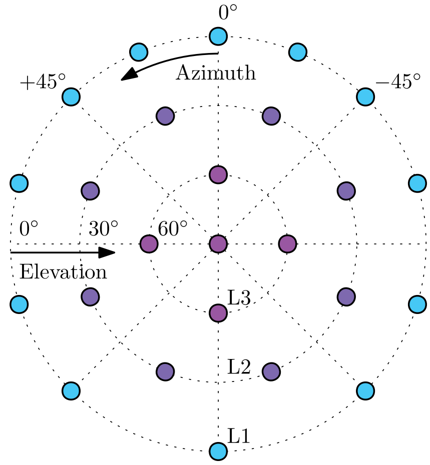


Fig. 1. Schematic diagram of a hemispherical loudspeaker setup with 25 channels (filled dots). The loudspeakers are grouped into three elevation layers: L1 (0° elevation), L2 (30° elevation), and L3 (≥60° elevation).

1 METHOD: SPATIAL GRANULAR SYNTHESIS

Granular synthesis is rooted in work by Dennis Gabor, who related time-frequency analysis with acoustical quanta to human perception of sound [35]. Early artistic work with spatialized, layered segments of sound is the piece "Concret PH" by Iannis Xenakis [36]. It was originally presented as an 11-channel tape piece, reproduced via 425 loudspeakers in the Philips Pavilion at Expo 58 [37]. Curtis Roads was the first to implement granular synthesis on digital computer platforms in the 1970s [38], and he explicitly mentions the potential of multichannel granular synthesis in his later works [39, 40]. Nuno Fonseca conceptualized and implemented particle systems for audio applications [41], in which particles can be complete audio files rather than the typically short audio segments used in granular synthesis (1–200 ms [40]). In more recent experimental work, spatialization of audio grains to a frontal array of loudspeakers was used to investigate perceived spatial extent in the horizontal and vertical dimension [42].

In spatial granular synthesis, each grain is assigned a position $\boldsymbol{x} = [x, y, z]^\top$ in space, which might equally be expressed in spherical coordinates $\boldsymbol{\Omega} = (\phi, \theta, r)$ [40, 27, 42, 26]. The possibilities of spatial grain distribution are manifold, e.g., one can aim for a directionally uniform distribution around the listener or restrict the distribution to a specific region in space. Several spatial audio techniques can be used to render the grains, e.g., amplitude-panning with loudspeaker arrays [43] or virtual sound source positioning with head-related transfer functions (HRTFs) [44, 45]. Ambisonics [46] allows encoding of grain objects positioned in a virtual environment, where reproduction takes place via binaural decoding to headphones [47], or via decoding to a multichannel loudspeaker system [48]. Lastly, discrete assignment of grains to the nearest available direction in an HRTF database or multichannel loudspeaker arrangement can serve as a baseline method for psychoacoustic experimentation, cf. Figs. 1 and 2.

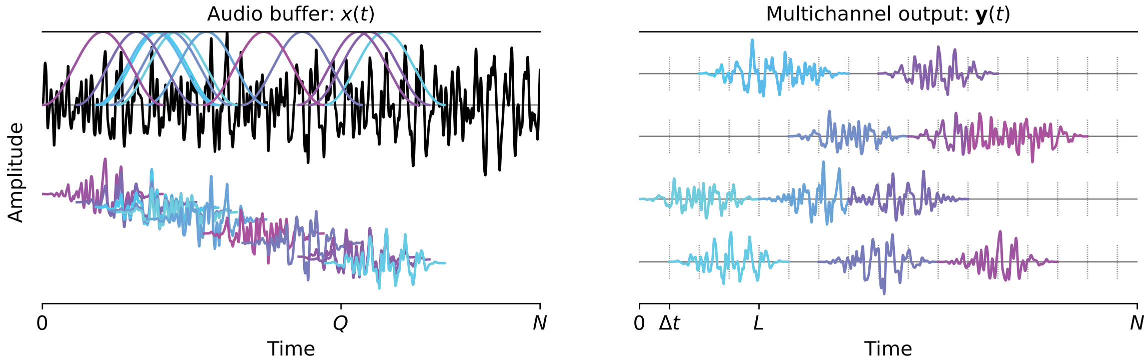


Fig. 2. Spatial granular synthesis extracts grains from an audio buffer and renders them to a multichannel output. The buffer is of length N seconds, and a grain of length L seconds is extracted at time index q_l from the buffer, where $q_l \sim U(0, Q)$ in this example. The synthesis period is constant at $\Delta t = L/4$ in this illustration.

1.1 Algorithm Definition

One may define a basic spatial granular synthesis algorithm with the following parameters:

- $\Delta t \rightarrow$ Time between spatialized grains,
- $L \rightarrow$ Grain length,
- $w \rightarrow$ Grain window/envelope,
- $Q \rightarrow$ Grain seed range in audio buffer, and
- $\boldsymbol{g} \rightarrow$ Weights for spatial rendering.

The algorithm requires access to a signal $x(t)$ of $N > L$ seconds of audio, which can be a recorded sample or a real-time input buffer:

$$x(t) = \begin{cases} x(t)\,, & \text{for } 0 \le t \le N \\ 0, & \text{else}\,. \end{cases} \tag{1}$$

The l-th grain is extracted at the buffer index q_l and a window function $w(t)$ is applied, to avoid artifacts in the output and shape its timbral properties. For stimulus generation in these experiments, a Hann window of length L seconds is used:

$$w(t) = \begin{cases} \sin^2\left(\frac{t\pi}{L}\right)\,, & \text{for } 0 \le t \le L \\ 0, & \text{else}\,. \end{cases} \tag{2}$$

Spatial rendering of grains to a J-channel output $\boldsymbol{y}(t) \in \mathbb{R}^J$ is achieved by multiplication with weights $\boldsymbol{g}_l(\boldsymbol{\Omega}_l) \in \mathbb{R}^J$, which are real-valued in case of discrete assignment, vector-base amplitude panning, and Ambisonics encoding [48]:

$$\boldsymbol{y}(t) = \frac{1}{\mathcal{G}} \sum_l \boldsymbol{g}_l \cdot w(t - \tau_l) \cdot x(t - \tau_l + q_l)\,, \tag{3}$$

where the summation considers active grains defined by $0 < (t - \tau_l) < L$ and $\tau_l = l\Delta t$ in case of strictly periodic synthesis. Convolution (*) with filter weights enables direct binaural synthesis using head-related impulse responses $\boldsymbol{g}_l(\boldsymbol{\Omega}_l,\, t) \in \mathbb{R}^2$ $(J = 2)$:

$$\boldsymbol{y}_{\mathrm{LR}}(t) = \frac{1}{\mathcal{G}} \sum_l \boldsymbol{g}_l * [w(t - \tau_l) \cdot x(t - \tau_l + q_l)]\,. \tag{4}$$

One can ensure constant loudness across varying grain densities by applying a gain factor $1/\mathcal{G}$ that compensates the spatio-temporal grain overlap $\Psi = L/\Delta t$ and the window function:

$$\mathcal{G} = \sqrt{\frac{L}{\Delta t}} \cdot \sqrt{\frac{1}{L} \int_0^L w^2(t)dt}\,, \tag{5}$$

assuming that extracted grains are uncorrelated. This is achieved by modulation or (uniform) random distribution of the extraction index $q_l \sim U(0, Q)$ with $Q \le (N - L)$.

The algorithm could in principle be applied to any kind of monophonic signal, be it a real-time input signal or an impulse response. A greater range Q for the grain extraction index $q_l \sim U(0,\, Q)$ enhances signal decorrelation in exchange for a longer response ("filter length") and signal displacement.

Formally, the spatial granular synthesis defined in Eq. (3) is equivalent to a time-variant FIR system, where the time variation of the FIR coefficients is dictated by the grain window (e.g., fade-in/out in case of a Hann window), and the delay of the taps varies randomly by q_l for every grain activation. The algorithm reduces to an LTI FIR system, if the time window is replaced by unity as $w(t) = 1$ and set $q_l = 0$.

The stimuli in this study were generated by an (offline) Python implementation of the algorithm defined in Eq. (3), which is openly available, see Sec. 4. Discrete (channel-based) grain assignment was used in the psychoacoustic experiments, to avoid any influence of order-dependent side-lobe levels using Ambisonics or the direction-dependent source widening of VBAP [48, 49]. In case of discrete assignment, the vector $\boldsymbol{g}_l$ has only one non-zero entry to assign a grain to a single target channel, cf. Fig. 2.

1.2 Instrumental Evaluation of Auditory Cues

By analyzing auditory cues such as the interaural coherence (IC), interaural time difference (ITD), and interaural level difference (ILD), and monaural spectral cues, one can assess perceptual differences between the synthesized stimuli and a stationary diffuse-field reference [14, 50]. This allows a more insightful interpretation of the experiment design and results presented in Sec. 2. Stimulus ear

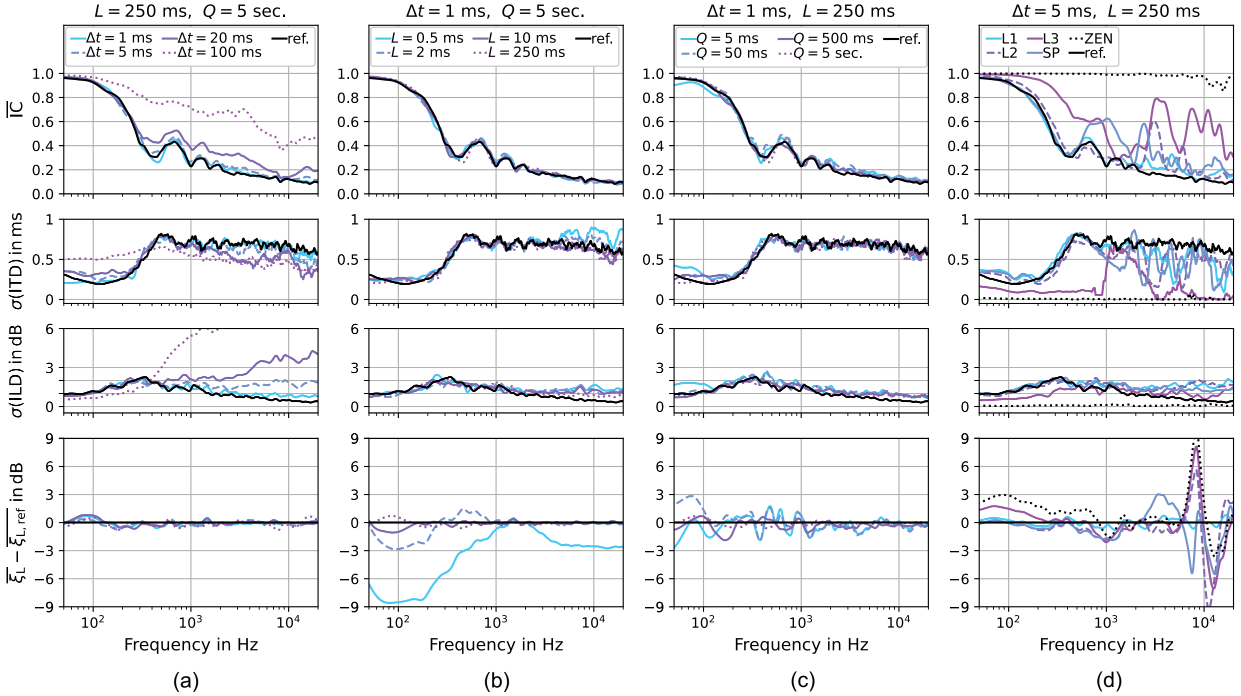


Fig. 3. Auditory cues for spatial granular synthesis versus a 2D diffuse-field reference (ref.; black curves). Each column varies a different synthesis parameter: time between grains Δt (a), grain lengths L (b), seed range Q (c), and directional assignment (d). Rows show different cues (top to bottom): Mean IC, standard deviation of ITD, standard deviation of ILD, and monaural difference between mean stimulus spectrum and mean diffuse-field spectrum.

signals $\boldsymbol{y}_{\mathrm{LR}}(t) = [y_{\mathrm{L}}(t),\ y_{\mathrm{R}}(t)]^{\top}$ are obtained by convolution of grain objects with (free-field) HRTFs, cf. Eq. (4). The auditory cues are computed in a time-dependent and frequency-dependent manner, and either temporal mean or standard deviation are presented (see APPENDIX A.1 for computation details).

The granular synthesis stimuli consist of pink noise grains spatialized via convolution with KU100 HRTFs [51]. Figs. 3(a)–3(c) show uniform random assignment in the horizontal plane (1° resolution), and Fig. 3(d) shows uniform random assignment within a direction subset available in the listening experiments (cf. Fig. 1). A 2D diffuse-field reference is simulated using a KU100 HRTF set of 360 directions (stationary, uncorrelated pink noise signals at 1° azimuth resolution), cf. black curves in Fig. 3.

In Fig. 3(a) the time interval Δt between spatialized grains is varied, whereas the grain length L and seed range Q are constant ($L = 250\,\mathrm{ms}$, $Q = 5\,\mathrm{s}$). At an interval $\Delta t = 100\,\mathrm{ms}$ (sparse condition), a high IC indicates that pronounced ITDs are extracted by the cross-correlation mechanism, suggesting that the individual sound events are well-localizable [20]. For an interval $\Delta t = 1\,\mathrm{ms}$, the synthesized sound field resembles a diffuse sound field, as the auditory cues show diffuse-field behavior (nearly identical to black curves). Low IC corresponds to high fluctuations of the ITDs, such that localization of individual sound events is impeded [52], causing a sensation of spaciousness and envelopment. In a diffuse sound field the magnitude of ILD fluctuation is limited to $\sigma(\mathrm{ILD}) \leq 2\,\mathrm{dB}$, meaning that short-time magnitudes of ILD are slightly above the just-noticeable difference (JND) of ILDs (≈0.5–1 dB) [53]. Compared to $\Delta t = 1\,\mathrm{ms}$, a spatialization at intervals $\Delta t \geq 5\,\mathrm{ms}$ causes pronounced ILDs above 1 kHz.

In Fig. 3(b) the effect of the window length L becomes apparent as a magnitude roll-off around the frequency $f = 1/L$, which can be seen for short grains of $L \leq 10\,\mathrm{ms}$. Note that grain lengths $L = 250\,\mathrm{ms}$ yield an output spectrum that corresponds to the reference condition (spectrum of the input signal).

In Fig. 3(c), the variable parameter is the seed range Q for the selection of grains from the audio buffer. Interestingly, for values of $Q \ll L$, in which phase-correlation between seeded grains tends to increase, the interaural cues still show diffuse-field behavior. However, spectral peaks and notches become visible in the magnitude spectra for $Q < L$, cf. Fig. 3(c), which resemble the comb-filtering behavior of (correlated) early reflections in rooms.

In Fig. 3(d), assignment of grains is evaluated for the direction subsets used in the listening experiments. The synthesis parameters are $\Delta t = 5\,\mathrm{ms}$, $L = 250\,\mathrm{ms}$, $Q = 5\,\mathrm{s}$, and directional assignment is uniformly random among the channels of a selected subset: L1, L2, L3, stereophonic (SP, ±45° azimuth) or monophonic zenith (ZEN, 90° elevation), cf. Fig. 1. The horizontal layer L1 yields an IC close to the 2D diffuse-field IC up to around 2 kHz with minor deviations at higher frequencies. The height layer L2 deviates more notably above 2 kHz, and L3 deviates clearly from the diffuse field even below 2 kHz, cf. top row of Fig. 3(d). The ZEN condition is a single-direction stimulus and produces highly correlated ear signals, such that fluctuations in ITD and ILD are absent. Spatialization to height layers yields pronounced spectral features above 6 kHz (pinna cues), cf.

bottom of Fig. 3(d). Most notably, a prominent energy peak is seen at 8 kHz, which is known to be a cue for sound source elevation, cf. Blauert's "directional bands" [54, 44].

2 LISTENING EXPERIMENTS

In this section, two listening experiments are presented. The first experiment investigates the effect of the temporal and directional density of sound events (grains) on envelopment, and the second experiment compares the perception of envelopment and engulfment using various height loudspeaker layers. The participants were given the following definitions of the attributes [4]:

- Envelopment (LEV): being surrounded by sound, and
- Engulfment (LEG): being covered by sound from above.

All participants completed the first experiment on envelopment without knowing that an additional attribute (engulfment) would be defined in the second experiment, in order to avoid bias in their ratings on envelopment regarding 2D vs. 3D ("with-height") conditions. Furthermore, none of the presented experiments made use of a reference condition, to avoid any assumptions on what conditions are most enveloping/engulfing.

2.1 Experiment I: Effect of Temporal and Directional Density on LEV

2.1.1 Setup and Design

The experiment was conducted at the IEM CUBE, an academic reproduction studio/venue with a reverberation time of $\mathrm{RT}_{30} = 0.5$ s. The hemispherical layout consists of 25 full-range, point-source loudspeakers by d&b audiotechnik and is shown in Fig. 1. The loudspeakers of the setup were individually equalized by 512-taps minimum-phase FIR filters to the mean loudspeaker response in third-octave bands, including frequency-independent gain factors that compensated the level differences as measured from the (double-octave smoothed) frequency responses.

During the experiment, the listeners were seated centrally. Their head orientation was not constrained, aiming for a natural listening situation as in a concert or installation.

The experiment used a multiple stimulus paradigm, however without a reference, in order to avoid predefining any type of sound field to be most enveloping. Each trial contained eight conditions of 2-s duration, designed to range from non-enveloping to potentially enveloping scenes. Participants rated the absolute, perceived envelopment of the eight stimuli on a continuous scale from 0 to 100 (0: not at all, 50: moderate, 100: full), presented via a graphical application on a laptop computer.

The stimuli were generated by the spatial granular synthesis algorithm described in Sec. 1, in which the algorithm extracts Hann-windowed grains from random positions in the audio input file and assigns them randomly (uniform distribution) to channels of a designated loudspeaker subset, cf. Fig. 1. The audio buffer was large enough ($N > 5$ s; $Q > 5$ s) to avoid any spectral effects of sampling critically short buffers, cf. bottom of Fig. 3(c).

The trials 1–4 used grains of length $L \in \{0.5, 250\}$ milliseconds extracted from a sound sample of either pink noise (trials 1+2) or a vocal quartet (EBU SQAM Track 48, trials 3+4). Within the trials, Δt was varied between $\Delta t \in \{100, 20, 5, 1\}$ milliseconds and the directional assignment was varied between 2D (L1, ear-height loudspeakers) and 3D (L1L2L3, hemisphere). The range of time intervals was chosen such that for $\Delta t = 100$ ms the sound events are perceived as localizable auditory events [20], whereas at $\Delta t = 1$ ms the stimuli approximate diffuse sound fields, cf. Sec. 1.2 and binaural auralizations (see Sec. 4). For the impulsive grains ($L = 0.5$ ms), no spatio-temporal overlap occurs, because even for the smallest $\Delta t = 1$ ms, $\Psi = L/\Delta t < 1$.

A fifth experimental trial was designed to vary only the directional density by restricting grain assignment to one of the following loudspeaker subsets: stereophonic (SP), quadraphonic (QP), 2D (L1), or 3D (combined layers L1L2L3). The loudspeaker signals of trial 5 were created by $L = 250$ ms Hann-windowed grains assigned randomly within the respective subset every $\Delta t = 1$ ms ($\Psi = 250$). Because of the high spatio-temporal grain overlap, the loudspeaker signals can be assumed to be approximately stationary noise signals in this trial. As a second independent variable in trial 5, the grains were extracted from either a pink noise or low-pass–filtered pink noise sample (12^{th}-order Butterworth with a cut-off frequency of 1.8 kHz).

Across all trials 1–5, the time between grains Δt was subject to controlled jitter, limited to 1% of Δt, in order to prevent signal periodicity. However, the inherent timbral effects of the window length are likely more relevant, cf. bottom of Fig. 3(b).

2.1.2 Results

Fifteen participants took part in the experiment, either staff or students of the authors' institution. The experimental results of trials 1–4 are shown in Fig. 4. Per trial, two independent variables were tested, namely the time Δt between spatialized grains and the type of spatialization (2D vs. 3D). To test the effect of Δt, pairwise Wilcoxon signed-rank tests between neighboring steps of Δt were conducted and report p values for the 2D and 3D spatialization variants are reported in Table 1. It turns out that most steps in Δt yield statistically significant differences in envelopment. For trial 4 (vocal grains of 250 ms length), the last step from $\Delta t = 5$ ms to $\Delta t = 1$ ms is significant for 3D spatialization, but not for 2D, in which ratings reach saturation for $\Delta t = 5$ ms.

To test the effect of 2D vs. 3D spatialization, pairwise Wilcoxon signed-rank tests were conducted, and p values are reported in Table 2. As seen in Table 2, the spatialization type does not lead to a significant difference for most conditions. However, in trial 4 at $\Delta t = 5$ ms, in which the 2D spatialization apparently reached saturation, the corresponding 3D condition was rated significantly lower ($p = 0.021$).

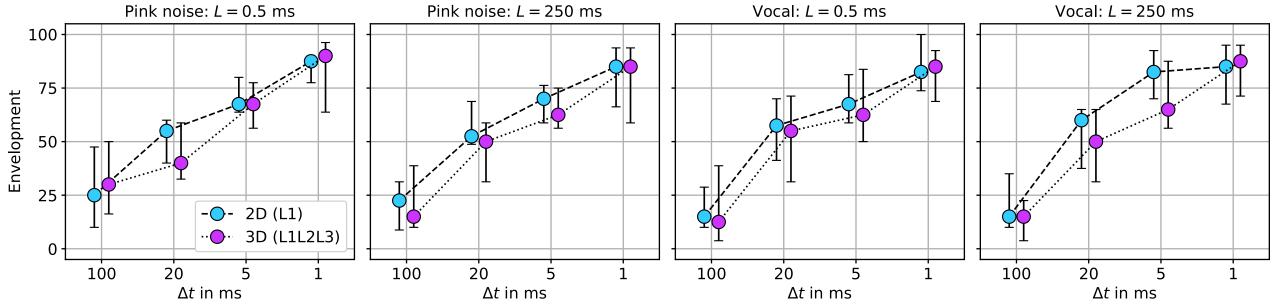


Fig. 4. Median responses and interquartile range (15 participants) for trials 1–4 of experiment I. For 2D conditions grains were randomly assigned to the ear-height loudspeakers (L1), whereas for the 3D conditions grains were assigned randomly among the total set of loudspeakers in the hemisphere (L1L2L3).

Table 1. Bonferroni-Holm corrected p values for three pairwise Wilcoxon signed-rank tests between Δt conditions of experiment I. Pink noise (PN) and vocal (VO) trials for each of the spatializations (2D/3D) and grains lengths (0.5 and 250 ms) are shown. Bold numbers indicate $p < 0.05$.

Δt (ms)	100 vs. 20	20 vs. 5	5 vs. 1
2D PN 0.5 ms	**0.005**	**0.007**	**0.024**
2D PN 250 ms	**0.006**	0.068	**0.009**
2D VO 0.5 ms	**0.000**	**0.024**	0.051
2D VO 250 ms	**0.004**	**0.003**	0.959
3D PN 0.5 ms	**0.013**	0.055	**0.024**
3D PN 250 ms	**0.008**	**0.041**	**0.033**
3D VO 0.5 ms	**0.016**	0.060	0.060
3D VO 250 ms	**0.000**	**0.004**	**0.013**

Table 2. Bonferroni-Holm corrected p values for four pairwise Wilcoxon signed-rank tests between 2D and 3D conditions of experiment I. Pink noise (PN) and vocal (VO) trials for each of the grains lengths (0.5 and 250 ms) are shown. Bold numbers indicate $p < 0.05$.

Δt	100 ms	20 ms	5 ms	1 ms
PN 0.5 ms	1.000	0.428	0.464	1.000
PN 250 ms	1.000	0.187	1.000	1.000
VO 0.5 ms	1.000	1.000	1.000	1.000
VO 250 ms	**0.028**	0.460	**0.021**	0.875

Table 3. Bonferroni-Holm corrected p values for three pairwise Wilcoxon signed-rank tests between SP, QP, horizontal layer L1, and hemispherical L1L2L3 conditions of trial 5 of experiment I. Bold numbers indicate $p < 0.05$.

	SP vs. QP	QP vs. L1	L1 vs. L1L2L3
Low-pass	**0.007**	0.944	0.530
Broadband	**0.003**	**0.013**	0.752

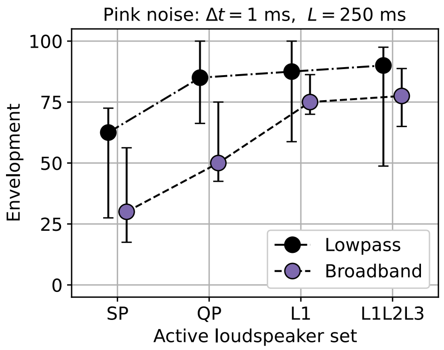


Fig. 5. Median and interquartile range for trial 5 of experiment I. The conditions correspond to SP (±45°), QP (±45° ± 135°), 2D ear-height layer (L1), and 3D hemisphere (L1L2L3).

The results of trial 5 are shown in Fig. 5. Wilcoxon signed-rank tests were conducted within the two signal groups (broadband and 1.8-kHz low-pass pink noise), cf. Table 3. A significant difference is found between SP and QP reproduction, for both broadband (unfiltered) and low-pass pink noise ($p = 0.003$, and $p = 0.007$). Interestingly, there is a significant difference between the QP and L1 conditions for broadband pink noise ($p = 0.013$), whereas there is no significant difference between those conditions for the 1.8-kHz low-pass pink noise ($p = 0.944$). Between the L1 (2D) and L1L2L3 (3D) conditions, no significant difference can be found, neither for broadband nor for low-pass pink noise signals.

2.1.3 Discussion

The results in Fig. 4 indicate that surrounding sound events at an interval of $\Delta t \leq 20$ ms evoke a moderate to high sensation of envelopment. In conditions without any spatio-temporal overlap of sounds ($L = 0.5$ ms), the perception becomes diffuse due to the processing lag of the human auditory system [17]. The perceptual integration time T must be greater than 20 ms, as a sensation of envelopment is formed for $\Delta t \leq 20$ ms. On the other hand, an upper bound for the integration time could be estimated as $T < 200$ ms, because envelopment ratings for $\Delta t = 100$ ms are low, as seen in Fig. 4, which suggests that individual sound events are perceived as localized and well-resolved auditory events. It is therefore conclusive to assume a perceptual integration time of $20\,\text{ms} < T < 200\,\text{ms}$, which is consistent with literature on the "binaural sluggishness" of the auditory system [17–20].

Auditory cues such as IC and ITD/ILD fluctuations can explain the ratings for the different intervals Δt, given the temporal window employed in the cue analysis corresponds to the perceptual integration time. At $\Delta t = 100$ ms, a high IC suggests localizability of successive sound events due to ITD cues, cf. Fig. 3(a). These ITD cues are less resolved for $\Delta t \leq 20$ ms, at which the IC decreases towards the diffuse-field IC. High-frequency (short-term) ILDs over 1 dB for $\Delta t = 20$ ms and $\Delta t = 5$ ms indicate noticeable spatial fluctuations, and might explain why ratings saturate at $\Delta t = 1$ ms, cf. Figs. 3(a) and 4.

Interestingly, the effect of 2D vs. 3D spatialization seems to be negligible, with a tendency that grain assignment to the 2D loudspeaker subset is more effective in producing envelopment. For a fixed Δt, the 3D conditions have the grains spread around the full hemisphere, leaving the horizontal layer with sparser signals. This could explain the trend towards lower ratings for 3D at $\Delta t = 20$ ms and $\Delta t = 5$ ms. At an interval of $\Delta t = 1$ ms, the ratings saturate for both 3D and 2D spatialization of grains.

The results of trial 5 agree with previous work on envelopment, which showed that for low-pass noise signals or reverberated music signals, four loudspeakers are perceptually close to a 24-loudspeaker (2D) reference [14, 6]. This is plausible because literature states that discrimination between a directionally sparse loudspeaker setup and a directionally dense reference is more difficult for low-pass noise signals [55]. Removing high-frequency signal content prohibits access to certain localization cues, especially high-frequency ILDs (and ITDs), cf. SP in Fig. 3(d), likely causing the reduced localizability and increased envelopment. Additionally, the results in Fig. 5 indicate that 2D spatialization (L1) is able to fully saturate perceived envelopment (even for broadband pink noise).

2.2 Experiment II: Effect of Height Loudspeakers and Signal Bandwidth on LEV vs. LEG

2.2.1 Setup and Design

The setup of the second experiment was equivalent to the first experiment, and all participants completed the second experiment after the first experiment. In addition to envelopment, a second attribute called engulfment was introduced to the participants. It is defined as "the sensation of being covered by sound from above", and the definition of envelopment was repeated as "the sensation of being surrounded by sound" [4]. The second experiment employs pink noise and an excerpt of the composition "Concret PH" by Iannis Xenakis. Although it would be possible to generate uncorrelated pink noise signals, the "Concret PH" excerpt requires a spatialization technique. Various techniques are thinkable [4], and this study used the spatial granular synthesis as presented above to spatialize both stimulus signals for uniformity.

The experiment was divided into two parts, according to the perceptual attributes envelopment and engulfment. In each part, participants rated only one of the attributes, and the order of the two parts was randomized. Each trial presented eight stimuli, either varying exclusively the active loudspeaker set (IIa) or varying both the signal bandwidth and the active loudspeaker set (IIb). The monophonic input to the spatial granular synthesis was either pink noise or an excerpt of the composition "Concret PH" by Iannis Xenakis. The length of the grains was $L = 250$ ms, spatialized at an interval $\Delta t = 5$ ms, both for the pink noise stimuli and the "Concret PH" stimuli. This gives a considerable spatio-temporal density of sound events, which allows for a moderate to high sensation of envelopment (and supposedly engulfment), cf. results of experiment I.

Table 4. Bonferroni-Holm corrected p values for two pairwise Wilcoxon signed-rank tests between layer conditions of experiment IIa. Bold numbers indicate $p < 0.05$.

	Envelopment		Engulfment	
	L2L3	L3	L2L3	L3
L1 (Pink)	0.169	**0.021**	**0.005**	**0.004**
L1 (Concret PH)	0.330	**0.005**	**0.009**	**0.026**

For the trials of type IIb, the stimuli were rated against their 1.8-kHz, low-pass–filtered versions (12th-order Butterworth). The loudspeaker sets defined for the experiment are the 0° elevation layer (L1), the 30° elevation layer (L2), and the group of remaining loudspeakers at 60° elevation plus the zenith loudspeaker (L3), cf. Fig. 1. A combination of the sets is denoted by concatenation of the abbreviations, e.g., L1L2 refers to the combined set of loudspeakers in the L1 and L2 layers. Lastly, two anchor stimuli were provided: an SP condition (±45° azimuth, 0° elevation) and a monophonic ZEN condition (90° elevation).

2.2.2 Results

Figs. 6 and 7 show results of the second listening experiment. Fig. 6(a) shows the effect of the active loudspeaker layer on envelopment and engulfment for the pink noise stimuli. The L1 condition (ear-height surround) obtained high ratings for envelopment, which is consistent with the results obtained in experiment I. When comparing L1 with the L2L3 and L3 conditions, one finds significantly lower envelopment ratings for the L3 condition ($p = 0.021$) but not for the L2L3 condition ($p = 0.169$), cf. Table 4 (left). The reported p values result from pairwise Wilcoxon signed-rank tests between L1 and the two other conditions. The results for the "Concret PH" stimuli in Fig. 6(b) show the same behavior, in which the envelopment rating of L1 is high, and ratings are significantly lower for L3 ($p = 0.005$).

Regarding engulfment, the L1 condition was rated low for pink noise stimuli, cf. Fig. 6(a), which is expected for a condition composed of horizontal-only, broadband sound. Whereas the rating of the L1L2 condition is higher than L1, engulfment further increased for conditions purely composed of the height layers L2 and L3, e.g., the difference between L1 and L3 shows to be highly significant in terms of engulfment ($p = 0.004$), cf. Table 4 (right). Notice that the monophonic zenith loudspeaker condition was rated lower than the horizontal L1 condition in terms of engulfment.

The results for the "Concret PH" stimuli show the same trends, cf. Fig. 6(b). The difference in engulfment between

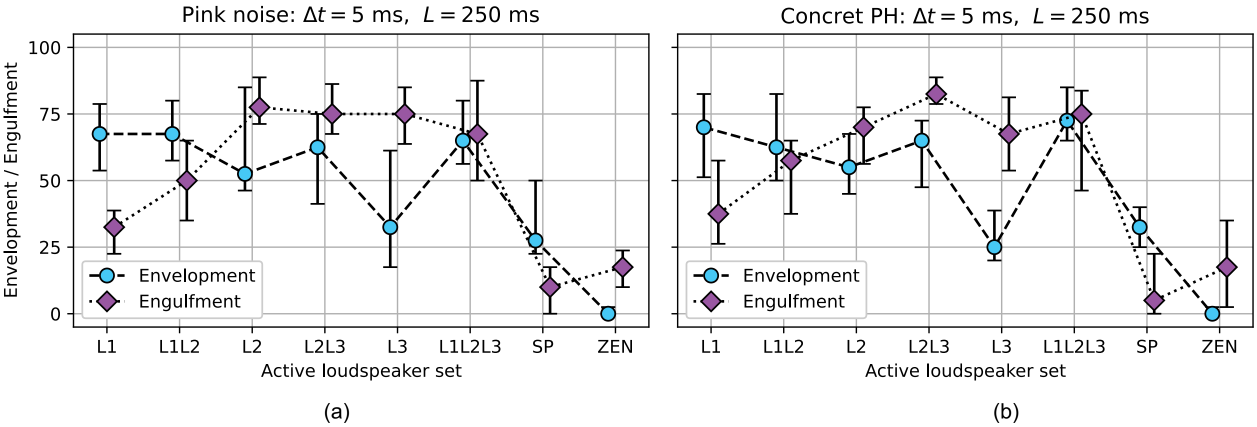
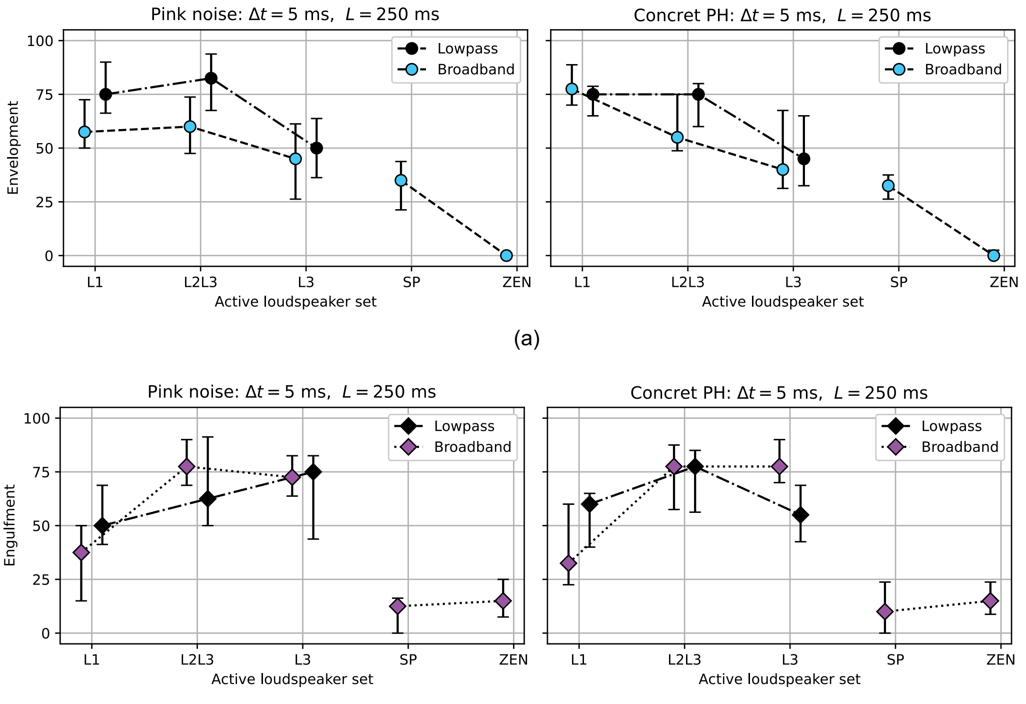


Fig. 6. Median and interquartile range (15 participants) for layer-based trials of experiment IIa.[*] Ratings for the pink noise stimuli are shown on the left (a) and ratings for the 'Concret PH' stimuli are shown on the right (b). Envelopment and engulfment was tested in separate trials, in which the independent variable was the loudspeaker subset: L1 (horizontal loudspeakers), L2 (loudspeakers at 30° elevation), and L3 (loudspeakers at ≥60° elevation). Hidden anchor conditions were included, namely an SP and a monophonic zenith condition (ZEN, 90° elevation).

L1 and L3 is also significant ($p = 0.026$), and the L2L3 condition achieved the highest rating regarding engulfment. Note that the L2L3 and L1L2L3 conditions achieved high ratings for envelopment and engulfment, for both pink noise and "Concret PH" stimuli.

The results in Fig. 7 show the effect of 1.8-kHz low-pass–filtered stimuli on envelopment and engulfment. Low-pass–filtered pink noise stimuli yield a relative increase in envelopment over broadband stimuli for the L1 and L2L3 conditions, cf. left image in Fig. 7(a). The reduction in envelopment from L1 to L3 is visible for all signal types and interestingly turns out to be significant for the low-pass–filtered pink noise stimuli ($p = 0.001$) and the low-pass–filtered "Concret PH" stimuli ($p = 0.001$), cf. Table 5 (left).

Regarding engulfment, ratings of the horizontal L1 conditions increase for low-pass–filtered stimuli, cf. Fig. 7(b). Certain height-layer conditions show reduced engulfment



(a)

(b)

Fig. 7. Median and interquartile range (15 participants) for layer-based and bandwidth-based trials of experiment IIb. Effect of loudspeaker set on perceived envelopment (a) and engulfment (b) for broadband (unfiltered) and low-pass–filtered stimuli (12th-order Butterworth with cut-off at 1.8 kHz). Tested layers are L1 (0° elevation), L2L3 (≥30° elevation), and L3 (≥60° elevation). Anchor conditions are SP and monophonic ZEN.

Table 5. Bonferroni-Holm corrected p values for two pairwise Wilcoxon signed-rank tests between layer conditions of experiment IIb: broadband (Bb.) and low-pass (Lp.) signals. Bold numbers indicate $p < 0.05$.

	Envelopment		Engulfment	
	L2L3	L3	L2L3	L3
L1 (Bb. Pink)	0.804	0.127	**0.001**	**0.002**
L1 (Lp. Pink)	0.400	**0.001**	0.136	0.151
L1 (Bb. Concret PH)	**0.026**	**0.011**	**0.009**	**0.010**
L1 (Lp. Concret PH)	0.656	**0.001**	**0.007**	0.679

ratings for the low-pass–filtered stimuli, such as L2L3 in case of the pink noise stimuli and L3 in case of the "Concret PH" stimuli. The difference in engulfment is significant between the L1 condition and the height conditions L2L3 ($p = 0.001$) and L3 ($p = 0.002$) for the broadband pink noise stimuli, whereas it is not significant for the low-pass–filtered pink noise stimuli, cf. Table 5 (right). Similarly, for the "Concret PH" stimuli the difference is significant between L1 and L3 for broadband (unfiltered) stimuli, but not so for the low-pass–filtered stimuli.

2.2.3 Discussion

The results of experiment II confirm that envelopment and engulfment are distinct perceptual attributes, as initially proposed by Sazdov et al. [32]. They are controllable by varying the active loudspeaker layer, which is especially clear when comparing the ratings for the L1 and L3 conditions. Although the L1 condition was rated as highly enveloping, it obtained low ratings for engulfment. Contrarily, the L3 condition was rated as engulfing but delivered a low sensation of envelopment. This is plausible, as the L3 condition did not supply any direct sound from directions below elevation $\theta = 60^\circ$, causing a lack of surrounding auditory events. However, the L3 condition provides height localization cues (8-kHz peak), and a sufficient amount of fluctuation in ITDs and ILDs, cf. Fig. 3(d), which seems to be the psychoacoustic foundation for engulfment. ZEN obtained low ratings for envelopment and engulfment, which is explained by the high IC ($\approx$1), corresponding to an absence of fluctuations in ITDs and ILDs. Engulfment clearly cannot be achieved by a single elevated sound source.

The second part of experiment II investigated the effect of the stimulus signal bandwidth. Regarding engulfment, low-pass–filtered stimuli reduced the difference between the horizontal L1 condition and the height conditions (L2L3 and L3), cf. Fig. 7(b). This could be due to the localization uncertainty (blur) introduced by the 1.8-kHz low-pass stimuli, which especially affects localization in the median plane, relying on monaural spectral cues above 2 kHz [56]. Although engulfment is controlled more stably with broadband stimuli, these results demonstrate that some low-pass–filtered stimuli were perceived as engulfing, which can be explained by binaural height localization cues available for laterally elevated sounds (lateral vertical planes) [57]. Butler and Humanski [57] showed that vertical localization with 3.0-kHz low-pass stimuli fails in the median plane, but is functional in a lateral vertical plane. These effects, together with dynamic listening cues available through head movements, could explain the ratings of engulfment for the low-pass–filtered stimuli, which are not as clearly separated as the ratings for broadband signals, cf. Fig. 7(b).

Regarding envelopment, both the 1.8 kHz low-pass–filtered stimuli and the broadband stimuli showed a degradation in ratings for the L3 layer compared to the L1 layer, which can be explained by the higher IC and decreased ITD fluctuations in a frequency region with high sensitivity to ITD cues (250–1000 Hz) [58], cf. Fig. 3(d).

3 CONCLUSION

This study proposed spatial granular synthesis as a method to generate sound fields with variable temporal and directional density. Listening experiments were conducted in a hemispherical loudspeaker array, and results indicate that listener envelopment requires surrounding sound events at intervals $\Delta t < 20$ ms. Reduction of the time interval between sound events showed a monotonic increase of listener envelopment in the experiments, in which saturation is reached for perceptually diffuse sound fields.

If multiple surrounding sound events occur within a sufficiently short time frame T, they cannot be individually resolved and localized. The auditory event becomes perceptually diffuse and enveloping, even when no simultaneous directional overlap was present. The perceptual integration time was found to be $20\,\text{ms} < T < 200\,\text{ms}$, which corresponds to literature on the binaural processing lag of the auditory system [17–20]. A running analysis of IC and ITD/ILD fluctuations could explain the experimental responses regarding temporal density, provided that the temporal analysis window is consistent with the perceptual integration time (e.g., $\mathcal{T} = 85$ ms).

Additionally, the design of the experiments did not make suggestions whether 3D (with-height) or 2D conditions would deliver a better sensation of envelopment and/or engulfment. This allowed to show that the ear-height loudspeaker layer contributes most effectively to envelopment, whereas height loudspeaker layers contribute primarily to engulfment. This can be explained by the fact that height layers provide monaural and binaural cues for vertical localization [57], but can lead to an increase in the IC depending on their elevation level.

The experiment results demonstrate a reduced control over engulfment for 1.8-kHz, low-pass–filtered stimuli. High-frequency signal content shows to be beneficial to control the sensation of engulfment, as it relies on perceptual cues for height localization, e.g., monaural spectral cues. In contrast, the sensation of envelopment can be enhanced by low-pass stimuli. The 1.8-kHz, low-pass–filtered pink noise increased envelopment especially for directionally sparse conditions (two or four active loudspeaker directions) due to the increased localization blur.

It should be noted that 3D (with-height) loudspeaker systems can render sound fields that will be perceived as both enveloping and engulfing (cf. L1L2L3 condition). This might be the cause why envelopment and engulfment were often understood as one sensation of "3D envelopment" or "subjective auditory diffuseness" [7, 15]. The results of this study underline that detailed investigations should treat both sensations separately.

4 OPEN DATA AND SOFTWARE

The authors provide open access to experiment data and code [59] and to binaural auralizations of the experiment stimuli [60]. Because the stimulus generation method was very effective and versatile as a tool, a real-time Ambisonic granular synthesis virtual studio technology plug-in was implemented and is available online at https://plugins.iem.at.

5 ACKNOWLEDGMENT

This work was partly funded by the Austrian Science Fund (FWF): P 35254-N.

6 REFERENCES

[1] S. Riedel and F. Zotter, "The Effect of Temporal and Directional Density on Perceived Envelopment," in *Proceedings of the 48th DAGA*, pp. 1059–1062 (Stuttgart, Germany) (2022 Mar.). https://pub.dega-akustik.de/DAGA_2022/data/articles/000089.pdf.

[2] J. S. Bradley and G. A. Soulodre, "The Influence of Late Arriving Energy on Spatial Impression," *J. Acoust. Soc. Am.*, vol. 97, no. 4, pp. 2263–2271 (1995 Apr.). https://doi.org/10.1121/1.411951.

[3] G. A. Soulodre, M. C. Lavoie, and S. G. Norcross, "Objective Measures of Listener Envelopment in Multichannel Surround Systems," *J. Audio Eng. Soc.*, vol. 51, no. 9, pp. 826–840 (2003 Sep.).

[4] H. Lynch and R. Sazdov, "A Perceptual Investigation Into Spatialization Techniques Used in Multichannel Electroacoustic Music for Envelopment and Engulfment," *Comput. Music J.*, vol. 41, no. 1, pp. 13–33 (2017 Mar.).

[5] J. Berg, "The Contrasting and Conflicting Definitions of Envelopment," presented at the *126th Convention of the Audio Engineering Society* (2009 May), paper 7808.

[6] S. Riedel and F. Zotter, "Surrounding Line Sources Optimally Reproduce Diffuse Envelopment at Off-Center Listening Positions," *JASA Express Lett.*, vol. 2, no. 9, paper 094404 (2022 Sep.). https://doi.org/10.1121/10.0014168.

[7] C. Eaton and H. Lee, "Subjective Evaluations of Three-Dimensional, Surround and Stereo Loudspeaker Reproductions Using Classical Music Recordings," *Acoust. Sci. Technol.*, vol. 43, no. 2, pp. 149–161 (2022 Mar.). https://doi.org/10.1250/ast.43.149.

[8] J. S. Bradley and G. A. Soulodre, "Objective Measures of Listener Envelopment," *J. Acoust. Soc. Am.*, vol. 98, no. 5, pp. 2590–2597 (1995 Nov.). https://doi.org/10.1121/1.413225.

[9] D. Griesinger, "The Psychoacoustics of Apparent Source Width, Spaciousness and Envelopment in Performance Spaces," *Acta Acoust. united Acoust.*, vol. 83, no. 4, pp. 721–731 (1997 Jul.).

[10] M. Morimoto and K. Iida, "A New Physical Measure for Psychological Evaluation of a Sound Field: Front/Back Energy Ratio as a Measure for Envelopment," *J. Acoust. Soc. Am.*, vol. 93, no. 4, pp. 2282–2282 (1993 Apr.). https://doi.org/10.1121/1.406551.

[11] H. Furuya, K. Fujimoto, C. Y. Ji, and N. Higa, "Arrival Direction of Late Sound and Listener Envelopment," *Appl. Acoust.*, vol. 62, no. 2, pp. 125–136 (2001 Feb.). https://doi.org/10.1016/S0003-682X(00)00052-9.

[12] J. Blauert and W. Lindemann, "Auditory Spaciousness: Some Further Psychoacoustic Analyses," *J. Acoust. Soc. Am.*, vol. 80, no. 2, pp. 533–542 (1986 Aug.). https://doi.org/10.1121/1.394048.

[13] T. Lokki, L. McLeod, and A. Kuusinen, "Perception of Loudness and Envelopment for Different Orchestral Dynamics," *J. Acoust. Soc. Am.*, vol. 148, no. 4, pp. 2137–2145 (2020 Oct.). https://doi.org/10.1121/10.0002101.

[14] K. Hiyama, S. Komiyama, and K. Hamasaki, "The Minimum Number of Loudspeakers and Its Arrangement for Reproducing the Spatial Impression of Diffuse Sound Field," presented at the *113th Convention of the Audio Engineering Society* (2002 Oct.), paper 5696.

[15] M. P. Cousins, F. M. Fazi, S. Bleeck, and F. Melchior, "Subjective Diffuseness in Layer-Based Loudspeaker Systems With Height," presented at the *139th Convention of the Audio Engineering Society* (2015 Oct.), paper 9427.

[16] M. Cousins, S. Bleeck, F. Fazi, et al., "The Effect of Inter-Channel Cross-Correlation Coefficient on Perceived Diffuseness," in *Proceedings of the 4th International Conference on Spatial Audio* (Graz, Austria) pp. 123–130 (2017 Sep.).

[17] J. Blauert, "On the Lag of Lateralization Caused by Interaural Time and Intensity Differences," *Audiology*, vol. 11, no. 5-6, pp. 265–270 (1972). https://doi.org/10.3109/00206097209072591.

[18] J. F. Culling and Q. Summerfield, "Measurements of the Binaural Temporal Window Using a Detection Task," *J. Acoust. Soc. Am.*, vol. 103, no. 6, pp. 3540–3553 (1998 Jun.). https://doi.org/10.1121/1.423061.

[19] D. W. Grantham and F. L. Wightman, "Detectability of a Pulsed Tone in the Presence of a Masker With Time-Varying Interaural Correlation," *J. Acoust. Soc. Am.*, vol. 65, no. 6, pp. 1509–1517 (1979 Jun.). https://doi.org/10.1121/1.382915.

[20] D. R. Perrott and S. Pacheco, "Minimum Audible Angle Thresholds for Broadband Noise as a Function of the Delay Between the Onset of the Lead and Lag Signals," *J. Acoust. Soc. Am.*, vol. 85, no. 6, pp. 2669–2672 (1989 Jun.). https://doi.org/10.1121/1.397764.

[21] S. J. Schlecht and E. A. Habets, "Time-Varying Feedback Matrices in Feedback Delay Networks and Their Application in Artificial Reverberation," *J. Acoust. Soc. Am.*, vol. 138, no. 3, pp. 1389–1398 (2015 Sep.). https://doi.org/10.1121/1.4928394.

[22] B. Alary, P. Massé, S. J. Schlecht, M. Noisternig, and V. Välimäki, "Perceptual Analysis of Directional Late Reverberation," *J. Acoust. Soc. Am.*, vol. 149, no. 5, pp. 3189–3199 (2021 May). https://doi.org/10.1121/10.0004770.

[23] E. Hoffbauer and M. Frank, "Four-Directional Ambisonic Spatial Decomposition Method With Reduced Temporal Artifacts," *J. Audio Eng. Soc.*, vol. 70, no. 12, pp. 1002–1014 (2022 Dec.). https://doi.org/10.17743/jaes.2022.0039.

[24] B. Alary, A. Politis, S. Schlecht, and V. Välimäki, "Directional Feedback Delay Network," *J. Audio Eng. Soc.*, vol. 67, no. 10, pp. 752–762 (2019 Oct.). https://doi.org/10.17743/jaes.2019.0026.

[25] V. Välimäki, B. Holm-Rasmussen, B. Alary, and H.-M. Lehtonen, "Late Reverberation Synthesis Using Filtered Velvet Noise," *Appl. Sci.*, vol. 7, no. 5, paper 483 (2017 May). https://doi.org/10.3390/app7050483.

[26] N. Barrett, "Spatio-Musical Composition Strategies," *Organ. Sound*, vol. 7, no. 3, pp. 313–323 (2002 Jun.). https://doi.org/10.1017/S1355771802003114.

[27] E. Deleflie and G. Schiemer, "Spatial Grains: Imbuing Granular Particles With Spatial-Domain Information," in *Proceedings of the Australasian Computer Music Conference*, paper 35 (Queensland, Australia) (2009 Nov.).

[28] C. Avendano and J.-M. Jot, "A Frequency-Domain Approach to Multichannel Upmix," *J. Audio Eng. Soc.*, vol. 52, no. 7/8, pp. 740–749 (2004 Jul.).

[29] C. Faller, "Multiple-Loudspeaker Playback of Stereo Signals," *J. Audio Eng. Soc.*, vol. 54, no. 11, pp. 1051–1064 (2006 Nov.).

[30] W. L. Martens and Y. Han, "Discrimination of Auditory Spatial Diffuseness Facilitated by Head Rolling While Listening to 'With-Height' versus 'Without-Height' Multichannel Loudspeaker Reproduction," in *Proceedings of the AES International Conference on Spatial Reproduction - Aesthetics and Science* (2018 Jul.), paper P4-3.

[31] H. Lee, "2D-to-3D Ambience Upmixing Based on Perceptual Band Allocation," *J. Audio Eng. Soc.*, vol. 63, no. 10, pp. 811–821 (2015 Oct.). https://doi.org/10.17743/jaes.2015.0075.

[32] R. Sazdov, G. Paine, and K. Stevens , "Perceptual Investigation into Envelopement, Spatial Clarity, and Engulfment in Reproduced Multi-Channel Audio," in *Proceedings of the 31st AES International Conference: New Directions in High Resolution Audio* (2007 Jun.), paper 25.

[33] N. Meyer-Kahlen, S. J. Schlecht, and T. Lokki, "Perceptual Roughness of Spatially Assigned Sparse Noise for Rendering Reverberation," *J. Acoust. Soc. Am.*, vol. 150, no. 5, pp. 3521–3531 (2021 Nov.). https://doi.org/10.1121/10.0007048.

[34] V. Välimäki, H.-M. Lehtonen, and M. Takanen, "A Perceptual Study on Velvet Noise and Its Variants at Different Pulse Densities," *IEEE Trans. Audio Speech Lang. Process.*, vol. 21, no. 7, pp. 1481–1488 (2013 Jul.). https://doi.org/10.1109/TASL.2013.2255281.

[35] D. Gabor, "Acoustical Quanta and the Theory of Hearing," *Nature*, vol. 159, no. 4044, pp. 591–594 (1947 May).

[36] I. Xenakis, *Formalized Music: Thought and Mathematics in Composition* (Pendragon Press, Hillsdale, NY, 1992).

[37] A. Valle, K. Tazelaar, and V. Lombardo, "In a Concrete Space: Reconstructing the Spatialization of Iannis Xenakis' *Concret PH* on a Multichannel Setup," in *Proceedings of the 7th Sound and Music Computing Conference* (Barcelona, Spain) (2010 Jul.).

[38] C. Roads, "Automated Granular Synthesis of Sounds," *Comput. Music J.*, vol. 2, no. 2, pp. 61–62 (1978).

[39] C. Roads, "Introduction to Granular Synthesis," *Comput. Music J.*, vol. 12, no. 2, pp. 11–13 (1988).

[40] C. Roads, *Microsound* (MIT Press, Cambridge, MA, 2004).

[41] N. Fonseca, "Particle Systems for Creating Highly Complex Sound Design Content," presented at the *137th Convention of the Audio Engineering Society* (2014 Oct.), paper 9132.

[42] M. Weger, G. Marentakis, and R. Höldrich, "Auditory Perception of Spatial Extent in the Horizontal and Vertical Plane," in *Proceedings of the 19th International Conference on Digital Audio Effects (DAFx)*, pp. 301–308 (Brno, Czech Republic) (2016 Sep.).

[43] V. Pulkki, "Virtual Sound Source Positioning Using Vector Base Amplitude Panning," *J. Audio Eng. Soc.*, vol. 45, no. 6, pp. 456–466 (1997 Jun.).

[44] J. Blauert, *Spatial Hearing: The Psychophysics of Human Sound Localization* (MIT Press, Cambridge, MA, 1997).

[45] D. R. Begault, E. M. Wenzel, and M. R. Anderson, "Direct Comparison of the Impact of Head Tracking, Reverberation, and Individualized Head-Related Transfer Functions on the Spatial Perception of a Virtual Speech Source," *J. Audio Eng. Soc.*, vol. 49, no. 10, pp. 904–916 (2001 Oct.).

[46] F. Zotter and M. Frank, *Ambisonics: A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality*, Springer Topics in Signal Processing, vol. 19 (Springer, Cham, Switzerland, 2019). https://doi.org/10.1007/978-3-030-17207-7.

[47] M. Zaunschirm, C. Schörkhuber, and R. Höldrich, "Binaural Rendering of Ambisonic Signals by Head-Related Impulse Response Time Alignment and a Diffuseness Constraint," *J. Acoust. Soc. Am.*, vol. 143, no. 6, pp. 3616–3627 (2018 Jun.). https://doi.org/10.1121/1.5040489.

[48] F. Zotter and M. Frank, "All-Round Ambisonic Panning and Decoding," *J. Audio Eng. Soc.*, vol. 60, no. 10, pp. 801–820 (2012 Oct.).

[49] M. Frank, "Source Width of Frontal Phantom Sources: Perception, Measurement, and Modeling," *Arch. Acoust.*, vol. 38, no. 3, pp. 311–319 (2013 Nov.). https://doi.org/10.2478/aoa-2013-0038.

[50] A. Walther and C. Faller, "Assessing Diffuse Sound Field Reproduction Capabilities of Multichannel Playback Systems," presented at the *130th Convention of the Audio Engineering Society* (2011 May), paper 8428.

[51] B. Bernschütz, "A Spherical Far Field HRIR/HRTF Compilation of the Neumann KU 100," in *Proceedings of the 39th DAGA*, pp. 592–595 (Meran, Italy) (2013 Mar.). https://doi.org/10.5281/zenodo.3928297.

[52] C. Faller and J. Merimaa, "Source Localization in Complex Listening Situations: Selection of Binau-

ral Cues Based on Interaural Coherence," *J. Acoust. Soc. Am.*, vol. 116, no. 5, pp. 3075–3089 (2004 Nov.). https://doi.org/10.1121/1.1791872.

[53] W. M. Hartmann and Z. A. Constan, "Interaural Level Differences and the Level-Meter Model," *J. Acoust. Soc. Am.*, vol. 112, no. 3, pp. 1037–1045 (2002 Sep.). https://doi.org/10.1121/1.1500759.

[54] V. G. Rajendran and H. Gamper, "Spectral Manipulation Improves Elevation Perception With Non-Individualized Head-Related Transfer Functions," *J. Acoust. Soc. Am.*, vol. 145, no. 3, pp. EL222–EL228 (2019 Mar.). https://doi.org/10.1121/1.5093641.

[55] O. Santala and V. Pulkki, "Directional Perception of Distributed Sound Sources," *J. Acoust. Soc. Am.*, vol. 129, no. 3, pp. 1522–1530 (2011 Mar.). https://doi.org/10.1121/1.3533727.

[56] R. Baumgartner, P. Majdak, and B. Laback, "Modeling Sound-Source Localization in Sagittal Planes for Human Listeners," *J. Acoust. Soc. Am.*, vol. 136, no. 2, pp. 791–802 (2014 Aug.). https://doi.org/10.1121/1.4887447.

[57] R. A. Butler and R. A. Humanski, "Localization of Sound in the Vertical Plane With and Without High-Frequency Spectral Cues," *Percept. Psychophys.*, vol. 51, no. 2, pp. 182–186 (1992 Mar.). https://doi.org/10.3758/BF03212242.

[58] A. Brughera, L. Dunai, and W. M. Hartmann, "Human Interaural Time Difference Thresholds for Sine Tones: The High-Frequency Limit," *J. Acoust. Soc. Am.*, vol. 133, no. 5, pp. 2839–2855 (2013 May). https://doi.org/10.1121/1.4795778.

[59] S. Riedel, "Data and Code Release," *Zenodo* (2023 Mar.). https://doi.org/10.5281/zenodo.7751713.

[60] S. Riedel, "Binaural auralizations of Listening Experiment Stimuli," *Zenodo* (2022 Oct.). https://doi.org/10.5281/zenodo.7342614.

[61] V. Hohmann, "Frequency Analysis and Synthesis Using a Gammatone Filterbank," *Acta Acust united Acust.*, vol. 88, no. 3, pp. 433–442 (2002 May/June).

[62] B. F. Katz and M. Noisternig, "A Comparative Study of Interaural Time Delay Estimation Methods," *J. Acoust. Soc. Am.*, vol. 135, no. 6, pp. 3530–3540 (2014 Jun.). https://doi.org/10.1121/1.4875714.

[63] B. R. Glasberg and B. C. Moore, "Derivation of Auditory Filter Shapes From Notched-Noise Data," *Hear. Res.*, vol. 47, no. 1-2, pp. 103–138 (1990 Aug.). https://doi.org/10.1016/0378-5955(90)90170-T.

A.1 COMPUTATION OF AUDITORY CUES

The presented analysis of auditory cues is time-dependent and frequency-dependent as in the auditory system ("running spectral analysis" [12]). The temporal window of the binaural hearing mechanism is modeled by analyzing sequential signal blocks of $\mathcal{T} = 85$ ms [18, 20]. The frequency resolution of the auditory periphery is modeled by a bank of (zero-phase) gammatone frequency windows $w_b(\omega)$ [61], in which b is the frequency band index and ω denotes the radial frequency. The short-time Fourier transforms of the ear signals are denoted as $Y_\mathrm{L}(\omega,\, t) = \mathcal{F}\{y_\mathrm{L}(t)\}_\mathcal{T}$ and $Y_\mathrm{R}(\omega,\, t) = \mathcal{F}\{y_\mathrm{R}(t)\}_\mathcal{T}$, in which $\mathcal{T}$ denotes the analysis block length. The following equations represent computations per signal block and therefore omit the notation of time dependence.

IC is computed as the normalized, maximum absolute value of the interaural cross-correlation function $R_\mathrm{LR}[b, \tau]$ [12]:

$$\mathrm{IC}[b] = \frac{\max\limits_{\tau} |R_\mathrm{LR}[b, \tau]|}{\sqrt{P_\mathrm{L}[b] \cdot P_\mathrm{R}[b]}}\,, \tag{6}$$

$$R_\mathrm{LR}[b, \tau] = \int_{-\infty}^{\infty} w_b^2(\omega) Y_\mathrm{L}^*(\omega) Y_\mathrm{R}(\omega)\, e^{j\omega\tau}\, d\omega\,, \tag{7}$$

$$P_\mathrm{L}[b] = \int_{-\infty}^{\infty} w_b^2(\omega) |Y_\mathrm{L}(\omega)|^2\, d\omega\,, \tag{8}$$

$$P_\mathrm{R}[b] = \int_{-\infty}^{\infty} w_b^2(\omega) |Y_\mathrm{R}(\omega)|^2\, d\omega\,, \tag{9}$$

where the search range for the lag τ is typically limited to $-1\ \mathrm{ms} \leq \tau \leq 1\ \mathrm{ms}$, and $(\,\cdot\,)^*$ denotes complex conjugation. The ITD and ILD are computed as [62]:

$$\mathrm{ITD}[b] = \operatorname*{argmax}_{\tau} R_\mathrm{LR}[b, \tau]\,, \tag{10}$$

$$\mathrm{ILD}[b] = 10 \cdot \log_{10}\left(\frac{P_\mathrm{L}[b]}{P_\mathrm{R}[b]}\right) \mathrm{dB}\,. \tag{11}$$

To investigate monaural ear signal spectra, the following is computed:

$$\xi_\mathrm{L}[b] = 10 \cdot \log_{10}(P_\mathrm{L}[b])\, \mathrm{dB}, \tag{12}$$

$$\xi_\mathrm{R}[b] = 10 \cdot \log_{10}(P_\mathrm{R}[b])\, \mathrm{dB}. \tag{13}$$

For the instrumental evaluation shown in Fig. 3, a pink noise buffer of $N = 10$ s duration is the input to the spatial granular synthesis. The output are binaural signals of 5 seconds duration, rendered using HRTFs of the Neumann KU100 dummy head [51]. The binaural signals are split into successive blocks of $\mathcal{T} = 85\,\mathrm{ms}$, and for each block, Eqs. (1)–(8) are evaluated for 320 gammatone magnitude windows $w_b(\omega)$ [61]. The band-pass windows are spaced on an equivalent rectangular bandwidth (ERB) frequency scale (1/8 ERB spacing), and each window covers one ERB [63].

Per simulated sound field stimulus, the authors compute the (temporal) mean $\overline{\mathrm{IC}}$ of the frequency-dependent interaural coherence. Because the interaural time and level differences are zero-mean for a balanced, left-right symmetric distribution of sound events, the standard deviations $\sigma(\mathrm{ITD})$ and $\sigma(\mathrm{ILD})$ are computed to measure the amount of temporal fluctuation [12]. Lastly, the difference between the mean left-ear spectrum of a stimulus $\overline{\xi_\mathrm{L}}$ and the mean left-ear spectrum of a 2D diffuse-field reference $\overline{\xi_\mathrm{L,ref}}$ is assessed..

THE AUTHORS

Stefan Riedel

Matthias Frank

Franz Zotter

Stefan Riedel is a researcher and Ph.D. student at the Institute of Electronic Music and Acoustics (IEM) of the University of Music and Performing Arts Graz (KUG). He graduated in electrical and audio engineering from TU Graz and KUG with a Master's thesis on mixed-order spherical beamforming in 2019. His current research focuses on the perception of envelopment in spatial sound reproduction. Further interests include binaural rendering for augmented/virtual reality, virtual acoustics, and psychoacoustics.

•

Matthias Frank is post-doctoral fellow at the Institute of Electronic Music and Acoustics (IEM) and deals with virtual acoustics, Ambisonics, musical acoustics, and psychoacoustics. He studied electrical and audio engineering at TU Graz and University of Music and Performing Arts (KUG) and graduated in 2009 with DI/M.Sc. In 2013, he finished his Ph.D. at KUG about the perception of auditory events created by multiple loudspeakers, and he is about to enter tenure track in 2023.

•

Franz Zotter is Associate Professor at the Institute of Electronic Music and Acoustics (IEM) and deals with virtual acoustics, Ambisonics, spherical beamforming, and sound reinforcement technologies. He graduated from TU Graz and University of Music and Performing Arts in Graz with a DI in electrical and audio engineering in 2004, was awarded a Ph.D. degree from University of Music and Performing Arts in 2009, got on tenure track in 2019, and was awarded tenure in 2023.