The Effect of Temporal and Directional Density on Perceived Envelopment

Stefan Riedel[1], Franz Zotter[1]

[1] *Institute of Electronic Music and Acoustics, University of Music and Performing Arts Graz, Austria, Email: riedel@iem.at*

Introduction

Listener envelopment (LEV) refers to the 'sensation of being surrounded by sound'. The perceptual attribute has been investigated in the fields of concert hall acoustics, spatial sound reproduction and electroacoustic music [1, 2]. In contrast to apparent source width (ASW), which refers to the extent of an auditory event, LEV is related to the overall immersive auditory quality. 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 [3]. 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 bandlimited noise signals or music stimuli [4]. An experiment on the directional perception of distributed sound sources confirmed that bandlimited signals reduce the perceptual difference between a sparse source arrangement and a dense reference [5].

The aforementioned experiments investigated the effect of the directional density in terms of the number of active loudspeakers, and their test signals were either stationary noise signals or reverberated music signals. In contrast, little to no knowledge seems to be available on the required temporal density of the reproduced sounds. This contribution presents a listening experiment that investigates variations in both the temporal and directional density of sound events. As a method for stimulus generation a spatial granular synthesis technique was implemented, as described below.

Experiment Setup and Design

The experiment was conducted at the 'IEM CUBE', an academic reproduction studio/venue with a reverberation time of RT30 = 0.5 s. The hemispherical layout consists of 25 loudspeakers and is shown in Fig. 1. The loudspeakers were individually equalized by 512-taps minimum-phase FIR filters to the mean loudspeaker response in third-octave bands, after application of broadband gain factors that compensated the volume differences as measured from the double-octave smoothed frequency responses. During the experiment the listeners were seated centrally. The 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 8 conditions of two seconds duration, designed to range from non-enveloping to potentially enveloping scenes. Participants rated the *'sensation of being surrounded by sound'* on a continuous scale from 0 to 100 (0: not at all, 50: moderate, 100: full).

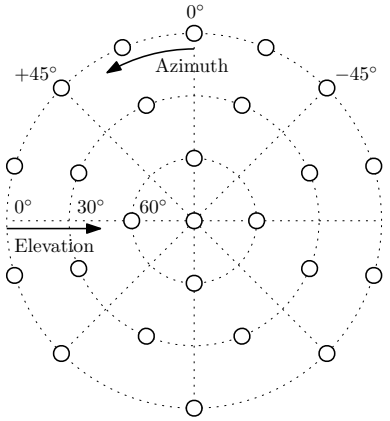

Figure 1: Experimental setup in the 'IEM CUBE' with 25 loudspeakers in a hemisphere arrangement.

The stimuli were generated by a spatial granular synthesis with the following parameters:

- Time Δt between spatialized wavelets
- Wavelet length L
- Directional assignment (2D/3D/subset)
- Source signal

The algorithm extracts Hann-windowed wavelets from random positions in the input file and assigns them to random loudspeaker channels (uniform random distribution) at time intervals Δt. The trials 1-4 used wavelets of length $L \in \{0.5, 250\}$ milliseconds sampled from pink noise (trial 1+2) or a vocal quartet sample (trial 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 (ear-height loudspeakers) and 3D (hemisphere). We can compute the effective wavelet overlap Ψ as

$$\Psi = \frac{L}{\Delta t}\,, \tag{1}$$

and observe that for the impulsive wavelets ($L = 0.5$ ms) no overlap occurs, as even for the smallest $\Delta t = 1$ ms we have $\Psi < 1$, cf. Tab. 1.

Ψ	100 ms	20 ms	5 ms	1 ms
0.5 ms	< 1	< 1	< 1	< 1
250 ms	2.5	12.5	50	250

Table 1: Overlap $\Psi = L/\Delta t$ for wavelets of length $L \in \{0.5, 250\}$ ms at intervals $\Delta t \in \{100, 20, 5, 1\}$ ms.

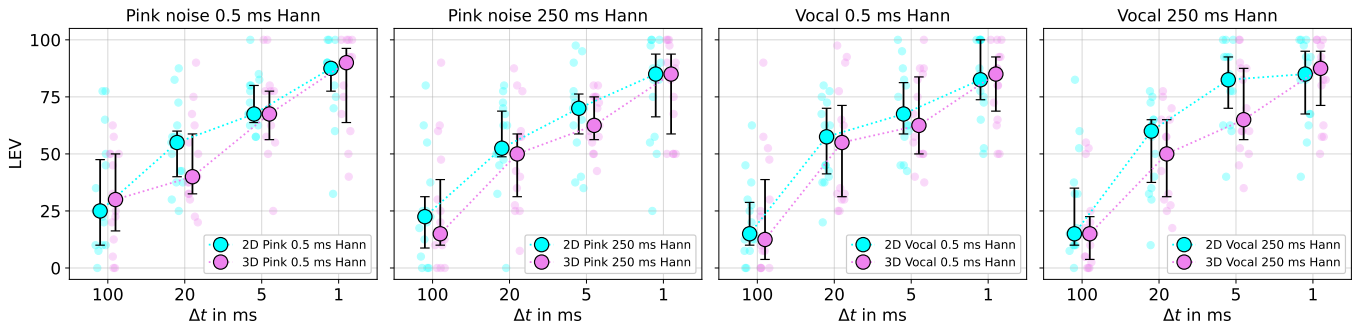

Figure 2: Median, interquartile range and individual responses of trials 1-4 (N = 15 participants). For 2D conditions (cyan) wavelets were randomly assigned to the ear-height loudspeakers only, whereas for 3D conditions (violet) wavelets were assigned randomly to the total set of loudspeakers in the hemisphere.

A fifth experimental trial was designed to vary the directional density by restricting the wavelet assignment to one of the following loudspeaker subsets: stereo ($\pm 45°$), quadraphonic ($\pm 45°, \pm 135°$), 2D ear-height, or 3D hemisphere. Accordingly, the number of active loudspeakers was $N_{LS} \in \{2, 4, 12, 25\}$. The loudspeaker signals were created by $L = 250$ ms Hann wavelets assigned randomly every $\Delta t = 1$ ms ($\Psi = 250$) to one of the channels of the respective subset. As a second independent variable, the wavelets were sampled from either pink noise or lowpass filtered pink noise (12^{th}-order Butterworth with a cut-off frequency of 1.8 kHz). Across all trials 1-5, the time between wavelets Δt was subject to controlled jitter, limited to 1% of Δt, in order to prevent signal periodicity.

Experiment Results

Trials 1-4

Fifteen participants took part in the experiment, either staff or students of the authors' institution. The experimental results of the trials 1-4 are shown in Fig. 2. The results show that for wavelets with $L = 0.5$ ms an interval of $\Delta t \leq 20$ ms is sufficient for a moderate to high sensation of envelopment. Even though the temporal and directional overlap was $\Psi < 1$ for conditions with $L = 0.5$ ms, the perception becomes diffuse due to the lag of localization. The perceptual integration time T must be greater than 20 ms, as a sensation of envelopment is formed for $\Delta t \leq 20$ ms. An upper bound for the integration time could be given as $T < 200$ ms. This is because the median ratings for $\Delta t = 100$ ms are low, which suggests the presence of highly localizable and well resolved auditory events rather than a perceived diffuseness. From the present experimental data, it is therefore conclusive to assume an integration time of 20 ms $< T <$ 200 ms. The physical overlap Ψ, cf. Tab. 1, does not appear to be suitable indicator of perception, which is shown by the range of the median ratings for $L = 0.5$ ms wavelet conditions which all give $\Psi < 1$.

Interestingly, the effect of 2D/3D loudspeaker subset seems to be negligible, with a tendency that wavelet assignment to the 2D loudspeaker subset seems to be more effective in producing envelopment when spatializing a limited number of wavelets. For a fixed Δt, the 3D conditions had the wavelets spread around the full hemisphere, leaving the horizontal layer with sparser signals. This could explain the trend towards lower ratings of 3D. It should *not* be concluded that height layers are of little use in spatial sound reproduction, however it seems that what they contribute is a distinct sensation referred to as 'engulfment' (sensation of being covered by sound [2]).

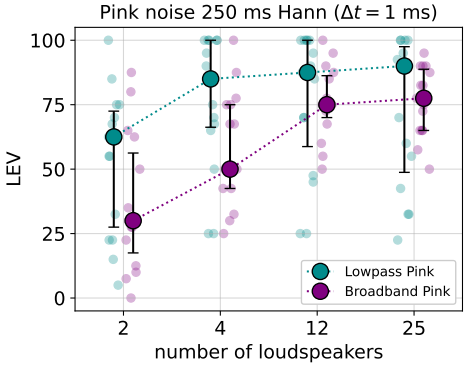

Figure 3: Median, interquartile range and individual responses for trial 5. The conditions correspond to stereo ($\pm 45°$), quadraphonic ($\pm 45°, \pm 135°$), 2D ear-height layer, and 3D hemisphere reproduction.

Trial 5

The results of trial 5 are shown in Fig. 3. A pairwise Wilcoxon signed-rank test with Bonferroni-Holm correction was conducted within the signal groups (broadband and lowpass). It shows that there is a significant difference between 4 and 12 loudspeakers for broadband pink noise ($p < 0.05$), while there is no significant difference between 4 and 12 loudspeakers for the 1.8 kHz lowpass pink noise. For both signal groups a significant difference between 2 and 4 loudspeakers could be found ($p < 0.05$). Between the 12 (2D) and 25 (3D) loudspeaker conditions, no significant difference can be found, neither for broadband nor for lowpass pink noise signals.

These results align with previous work on spatial impression, which showed that for bandlimited noise or reverberated music signals, 4 loudspeakers are perceptually close to a 24 loudspeaker (2D) reference [4]. Experiments on the perception of distributed sound sources showed that the discrimination between a sparse stimulus setup and a dense reference setup was more difficult for bandlimited signals [5].

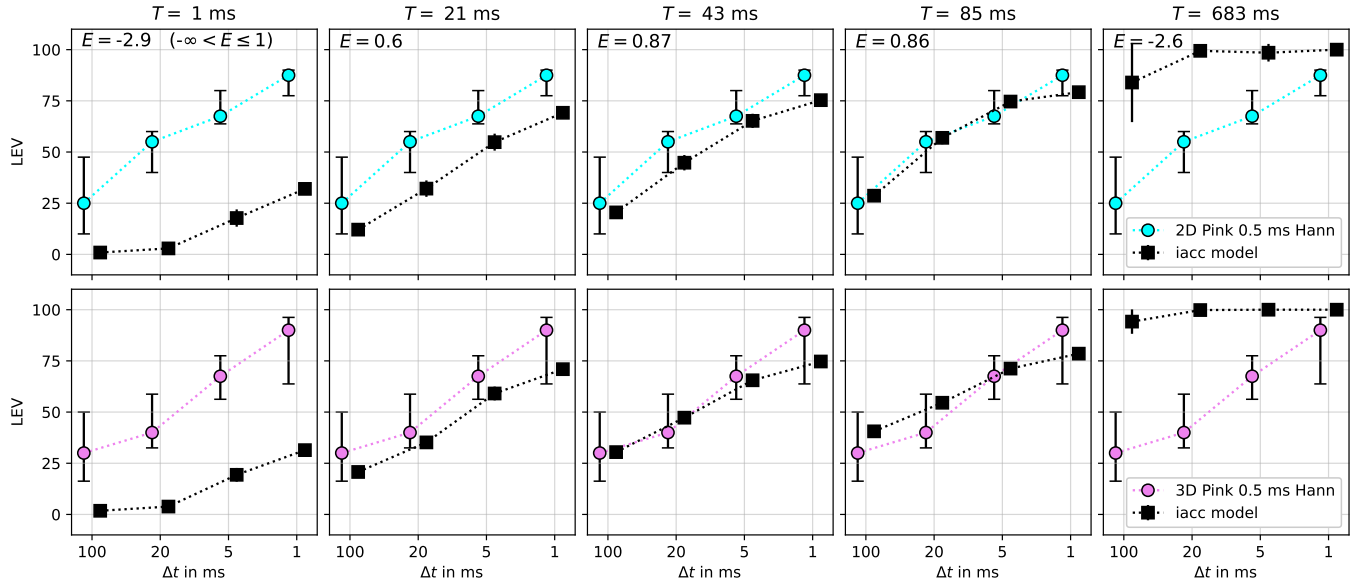

Figure 4: IACC-model vs. data (trial 1) for various integration times T. The Nash-Sutcliffe coefficient E as a measure for the 'goodness-of-fit' is computed per integration time T (per column), showing good agreement for $T = 43$ ms and $T = 85$ ms.

LEV model based on IACC

Model Definition

In this section we present a simple model for LEV based on the interaural cross-correlation coefficient (IACC). The idea is to find an integration time parameter $T = T_{\text{iacc}}$ that optimally fits the experiment data, and to verify whether it lies in the abovementioned range ($20\,\text{ms} < T < 200\,\text{ms}$). The model computes the IACC for each signal block of the length T and maps it to a LEV prediction. The discrete time index $n \in \mathbb{N}_0$ refers to the time-varying estimate

$$\text{LEV}[n] = 1 - \frac{\max(\,\text{IACC}[n] - \text{IACC}_{\text{ref}}, 0\,)}{1 - \text{IACC}_{\text{ref}}}\,, \qquad (2)$$

$$\text{IACC}[n] = \max_{\tau} \left| \frac{\int_{nT}^{nT+T} x_{\text{L}}(t) \cdot x_{\text{R}}(t+\tau)dt}{\sqrt{\int_{nT}^{nT+T} x_{\text{L}}^2(t)dt \cdot \int_{nT}^{nT+T} x_{\text{R}}^2(t)dt}} \right|\,, \qquad (3)$$

where the search range for the lag τ is limited to $-1\,\text{ms} \leq \tau \leq 1\,\text{ms}$. The reference IACC_{ref} is computed as the diffuse-field $\text{IACC}_{\text{diff}}(T)$ plus a perceptual threshold ϵ:

$$\text{IACC}_{\text{ref}} = \text{IACC}_{\text{diff}}(T) + \epsilon\,, \qquad (4)$$

where ϵ was set to $\epsilon = 0.2$ in this study. The diffuse-field $\text{IACC}_{\text{diff}}(T)$ and the perceptual threshold ϵ are used to calibrate the model. Note that the diffuse-field $\text{IACC}_{\text{diff}}$ may generally be larger than zero. The reasons are the coherence of the ear signals at low frequencies and the finite integration time T (block size). In our study $\text{IACC}_{\text{diff}}(T)$ was computed from diffuse-field ear signals simulated using a spherical HRTF set of the KU100 dummy head (TH Koeln set). The time-varying estimate $\text{LEV}[n]$ is finally averaged over time with energy weights

$$\text{LEV} = \frac{1}{\text{N}} \sum_{n} w[n] \cdot \text{LEV[n]}\,, \qquad (5)$$

where the weights $0 \leq w[n] \leq 1$ are normalized to the maximum composite RMS of the binaural stimulus signal. The dynamic range was set to 20 dB, such that signal blocks with a relative RMS of -20 dB and below are assumed to have no perceptual relevance. Finally, we can compute mean and standard deviation of this estimator across different head orientations. In this study, we used the following orientations: $\{-90, -60, -30, 0, +30, +60, +90\}$ degree azimuth.

The binaural stimuli were created by convolution of the loudspeaker signals with binaural room impulse responses (BRIRs). The BRIRs were captured with a Neumann KU100 dummy head located at the center of the loudspeaker arrangement, which corresponds to the position of the participants during the experiment. The measured BRIRs were not significantly shortened and entailed direct sound, early reflections and reverberation of the experimental room (BRIR length of RT30 = 0.5 s).

Model Evaluation

Figures 4 and 5 show evaluations of the model for various integration times T. The effect of the integration time can be seen clearly in Fig. 4: for a very short integration time $T = 1$ ms, the model would underestimate the perceived envelopment. On the other hand, if we used an integration time $T = 683$ ms, even perceptually sparse and localizable conditions would yield high LEV estimates. The Nash-Sutcliffe coefficient is used to measure the 'goodness-of-fit' between the data points d_i (median ratings) and model predictions m_i:

$$E = 1 - \frac{\sum_i (d_i - m_i)^2}{\sum_i (d_i - \overline{d})^2}\,, \qquad (6)$$

where $\overline{d}$ refers to the mean of the data points. The coefficient generally takes values in the range $-\infty < E \leq 1$. It is maximized for integration times $43\,\text{ms} \leq T \leq 85\,\text{ms}$ for the data of trial 1, reaching a value of $E = 0.87$.

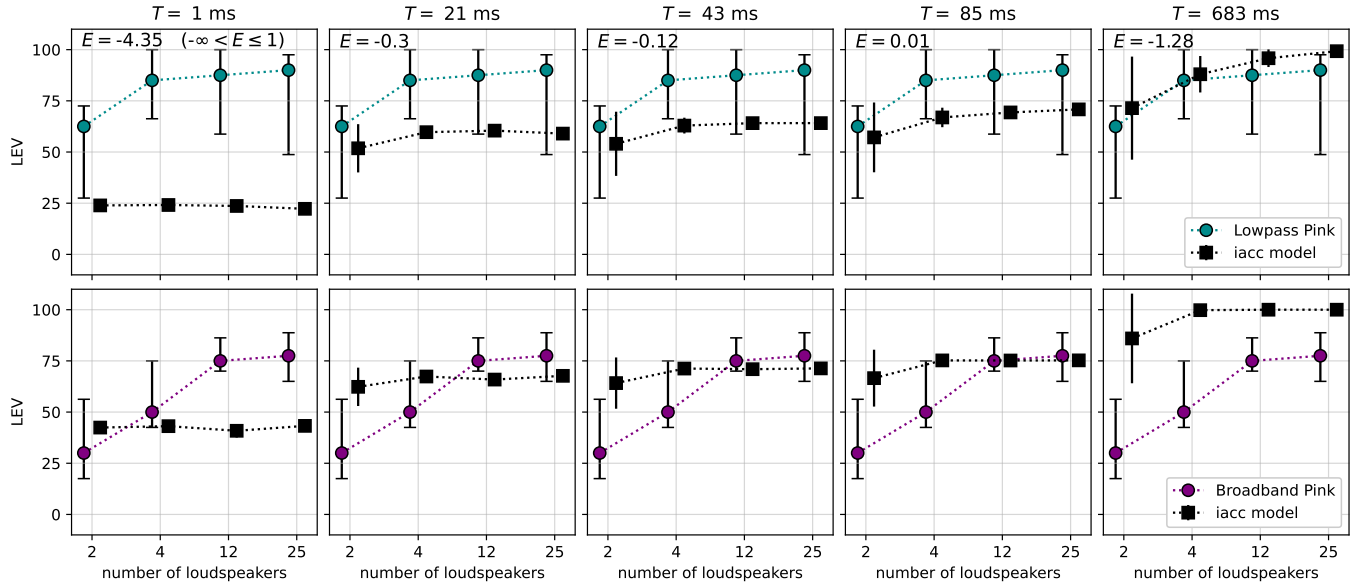

Figure 5: IACC-model vs. data of trial 5 for various integration times T. The Nash-Sutcliffe coefficient E is computed per integration time T (per column), showing that the IACC model cannot explain the bandwidth dependency.

This corresponds well with the estimation of T from the experimental results themselves ($20\,\text{ms} < T < 200\,\text{ms}$). Unfortunately, the IACC model does not explain the bandwidth dependency found in trial 5, cf. Fig. 5. A model that can explain the increased localizability of the broadband signals would therefore be desirable. While functional models for spectral localization cues [6] seem promising under anechoic conditions, they are overly fragile under reverberant conditions, when the spectrum of the ear signals is smeared by the reverberation. An artificial neural network that replicates the auditory system might be able to robustly recognize and match patterns from noisy real-world ear signals [7]. Alternatively, parametric vector models that assume a priori knowledge of the sound field might be a viable pathway towards robust predictive models for LEV.

Conclusion

We investigated how the temporal and directional density of sound events affects the perception of listener envelopment. If multiple sound events occur in a short time frame T, they cannot be individually resolved any more and the auditory event becomes perceptually diffuse, even when no simultaneous directional overlap was present. The perceptual integration time was found to be $20\,\text{ms} < T < 200\,\text{ms}$. This was confirmed by a model based on interaural cross-correlation, which shows optimal predictions for integration times of $43\,\text{ms} \leq T_{\text{iacc}} \leq 85\,\text{ms}$. Additionally, the design of the experiment did not make suggestions of whether 2D or 3D conditions would deliver a higher envelopment, and this allowed us to show that the 2D (ear-height) loudspeaker layer contributes most substantially to envelopment.

Regarding the effect of directional density, our experiment suggests that broadband pink noise signals require a dense directional coverage to elicit envelopment, while 1.8 kHz lowpass filtered pink noise can be enveloping even with just 4 active loudspeaker directions, assuming a central listening position ('sweet-spot'). We explain this by the higher localizability of broadband sound sources, which is in agreement with previous work on the bandwidth-dependent discrimination of distributed sound sources [4, 5].

Appendix

Binaural auralizations of the stimuli can be found at `https://phaidra.kug.ac.at/o:126272` or `https://cloud.iem.at/index.php/s/Z2G4XW2nn4M5AAM`.

References

[1] G. A. Soulodre, M. C. Lavoie, and S. G. Norcross, "Objective measures of listener envelopment in multichannel surround systems," *JAES*, vol. 51, no. 9, pp. 826–840, 2003.

[2] H. Lynch and R. Sazdov, "A perceptual investigation into spatialization techniques used in multichannel electroacoustic music for envelopment and engulfment," *CMJ*, vol. 41, no. 1, pp. 13–33, 2017.

[3] J. S. Bradley and G. A. Soulodre, "Objective measures of listener envelopment," *JASA*, vol. 98, no. 5, pp. 2590–2597, 1995.

[4] K. Hiyama, S. Komiyama, and K. Hamasaki, "The minimum number of loudspeakers and its arrangement for reproducing the spatial impression of diffuse sound field," in *113th AES Conv.*, 2002, paper 5696.

[5] O. Santala and V. Pulkki, "Directional perception of distributed sound sources," *JASA*, vol. 129, no. 3, pp. 1522–1530, 2011.

[6] R. Baumgartner, P. Majdak, and B. Laback, "Modeling sound-source localization in sagittal planes for human listeners," *JASA*, vol. 136, no. 2, pp. 791–802, 2014.

[7] A. Francl and J. H. McDermott, "Deep neural network models of sound localization reveal how perception is adapted to real-world environments," *Nature Human Behaviour*, vol. 6, no. 1, pp. 111–133, 2022.