Audio Engineering Society

Convention Paper 10251

Presented at the AES 159th Convention
2025 October 23–25, Long Beach, CA, USA

This paper was peer-reviewed as a complete manuscript for presentation at this convention. This paper is available in the AES E-Library (http://www.aes.org/e-lib), all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

Differentiating Sensations of Sound Envelopment in Spatial Sound Synthesis Approaches: an Explorative Study

Timothy Schmele[1,2] and Eleonora De Filippi[3]

[1]*Eurecat, Centre Tecnològic de Catalunya, Multimedia Technologies Group, Barcelona, Spain*
[2]*University of Music Karlsruhe, Institute for Musicology and Music Informatics, Karlsruhe, Germany*
[3]*ICube Laboratoire, SDC Data Science and Knowledge Team, University of Straßbourg, France*

Correspondence should be addressed to Timothy Schmele (tim.schmele@eurecat.org)

ABSTRACT

Investigating the enveloping effect of room acoustics has a long history, while research on the perceptual effects related to creative spatialization methods, often used in electroacoustic music, remains relatively scarce. These include spectral methods that distribute each frequency bin into a separate location in space, or temporal methods, where each grain of audio is given an individual position. Recently, a technique to synthesize sound spectra using nothing but spatial transformations has been presented that claims to decorrelate the sound spatially using high angular velocities. Backed by an extensive literature review, this study invited 18 expert listeners to evaluate these creative spatialization approaches on a horizontal 8 channel circular array across 6 perceptual dimensions related to envelopment using 3 different types of source sounds. In particular, it investigated if different spatial sound synthesis techniques can be considered perceptually unique in their spatialization approach. The results are significant, showing strong effects across all dimensions for all approaches chosen. Pairwise comparison also indicates that each creative spatialization technique might be indeed unique in its approach to create spatial impressions, but future work and standardization will be necessary.

1 Introduction

Sound envelopment is a type of spatial auditory impression that is often simply defined as the feeling of being surrounded by sound [1, 2, 3]. While research on this topic is mainly associated with the evaluation of concert hall acoustics [2, 3, 4, 5, 6, 7, 8], recent developments in the field of sound synthesis techniques that are combined with spatialization approaches are broadening the perspective this auditory percept [9, 10, 11]. As simple as the initial definition above may be, today it is understood that the sensation of being surrounded by sound does not always result in the same feeling in each listening situation [6, 5, 12]. The definitions for envelopment vary widely throughout literature [6] and a more contemporary understanding of envelopment is that of a multidimensional percept [4, 5, 13], similar to timbre [14]. That subjective assessments of concert halls need to be evaluated with more than one perceptual dimension has indeed already been postulated since the latter mid 20$^{\text{th}}$ century [15, 16]. However, only later was listener envelopment [17, 18] separated from the apparent source width [1, 18] as a distinct auditory percept. Since then, our understanding of envelopment has broken most of these terms down further into more component attributes [5, 6, 16, 19]. For example, it

has been shown that not only can listeners distinguish between source width gradients, but are also able to distinguish between different auditory shapes created using a 2D grid of decorrelated loudspeakers in varying configurations [20].

Thus, not all auditory spatial sensations of, or related to, sound envelopment are created equal. Composers of electroacoustic music in particular, focusing mainly on the loudspeaker as a sound producing instrument [21], have been practicing creative ways to spatialize sound over loudspeaker arrays since the mid of the 20th century [9, 12]. These can span from manually driving gain faders [22, 23] to intricate algorithms [9, 24]. A subset of these approaches are sometimes referred to as spatial sound synthesis, since they intend to create novel spatialization approaches by intertwining sound synthesis with spatialization [11, 25, 26, 27].

While these approaches may be valid within the context of artistic research, in many cases it remains unclear how effective they are from the perspective of the perceptive listener, with limited studies done in this field [10, 12] and the necessary vocabulary to put the spatial sensations into words may also still be missing [6]. The intention of this study is to further the research in this field. In particular, this study investigates if spatial sound synthesis approaches demonstrate the potential to create novel sensations of envelopment, by selecting suitable perceptual descriptors as potential perceptual dimensions of envelopment and comparing them against each other. Hence, the present study is coined explorative precisely because of the high degree of uncertainty in terminology and method.

1.1 The Perceptual Dimensions of Envelopment

Despite the lack of consensus on what constitutes the sensation of envelopment in detail, it is an important aspect of our auditory experience that requires further research [6], as also exemplified by the large research community formed around the evaluation of concert hall acoustics. For example, unlike the definition cited above [1, 2, 3], both Lindau et al. [7] and Weinzierl et al. [8] explicitly constrain envelopment to being a "[s]ensation of being spatially surrounded by the reverberation" [7] and relate an increased sense of envelopment to a decrease in localization accuracy. Rumsey [5] in particular laments the lack of rigorous investigation and clarity in standard definitions of what he calls *spatial attributes*. He stresses that to identify a spatial attribute, "[t]hey should enable meaningful and sensitive distinctions to be made between the products or techniques under test, and they should enable repeatable judgments." [5] Ideally, these identified attributes should be unambiguous and preferably unidimensional.

Rumsey [5] provides an extensive review on the many terms used to describe auditory spatial attributes, including many that relate to envelopment. Berg [6] tackles the terminology issue directly by analyzing the scientific corpus of the AES E-Library for terms specifically related to envelopment. Lindau et al. [7] and [8] employ a Focus Group approach to identify key terminology and develop the Spatial Audio Quality Inventory (SAQI) and Room Acoustical Quality Inventory (RAQI) respectively, evaluating interviews and discussion rounds with experts in the field. In Hawkes and Douglas [16] we can also find an exhaustive summary of the study performed by Beranek [15], providing a more historical view on the terms brought forth to categorize the subjective effects of concert hall acoustics. Other lists that were consulted for this study were the terms brought forth by Toole [19], in an effort to standardize listening tests, and the literature review done by Spors et al. [13], for the purpose of evaluating the quality of loudspeaker-based spatial sound reproduction techniques, among others.

Immersion is most likely the term most closely related to envelopment and is often used interchangeably. Berg [6] concludes, however, that immersion is a more complex term, where audition is often combined with other modes and is mostly used in the virtual reality domain. He defines immersion as a system's ability to create a sense of presence, naturalism and realism [6]. In this context, the ubiquitous term *fidelity* is also used, applied to the a system's ability to agree with a reference [13]. This is sometimes also described as *naturalness* [7]. *Engulfment*, another closely related term, is said to distinguish itself from envelopment in that it applies to the sensation of being covered by sound, mainly through the used of overhead loudspeakers. [12, 28]

Spaciousness is another blanket term used to cover both source width and envelopment [17, 29], also defined as "[...] a characteristic spatial spreading of the auditory events [...]" [2], but is sometimes used more specifically to only denote large reverberant spaces [5, 13]. Related to room acoustics, it is both linked to early lateral reflections [4] and the characteristics of the late reverberation tail [29]. Other synonyms can be *spatial*

impression, *spatial responsiveness*, *subjective diffusion* and also *ambiance* [2].

While *width* has previously been used interchangeably with envelopment, it is now understood as a percept related to the source itself, whereas envelopment relates to the overall sound scene surrounding the listener [1, 5, 13, 17, 19] or the width of a group of sources [19]. However, while Rumsey [5] separates width from envelopment, he attributes both a width and envelopment to each the source, ensemble and environment, similarly to what is described in Berg [6].

Presence, the sense of being there, or inside an enclosed space [7], is also sometimes considered equivalent to spatial impression [5]. It is also said to be equivalent to proximity, intimacy and engagement [3]. *Intimacy*, for example, is sometimes described a sense of closeness [18], which relates to the sense being in an enclosed space, but also a sense of involvement with a performance [18], which relates to the aforementioned engagement. *Engagement*, in turn, can have an effect on *impact*, determining the interest or effectiveness of a sound scene [3].

Intimacy is also said to arise from the *proximity* to the sound source [3, 16], which relates to distance. *Distance* is an auditory percept that is generally associated with the distance of sound sources to the listener [5, 19], but can also sometimes be applied to sound events in general [13]. As such, it relates to the sense *depth*, which is a sense of perspective in a scene [5, 13]. *Perspective* itself has been related to presence once more, stating that a reduction of perspective gives the impression as if listening through an opening onto the sound scene, in turn, increasing separation from the scene and reducing envelopment [19].

There are also several terms to denote distortions of the enveloping percept in both space and time. *Stability* relates to the temporal domain and indicates to which degree sources or the entire scene remains stable or changes over time [5, 13]. Conversely, there is *robustness*, which denotes the degree of change of an auditory event in relation to the listeners movement [13]. *Skew* captures the sense of a sound scene being skewed [5], *disturbance* signifies the presence of sufficiently loud and delayed lateral reflections [30] and *homogeneity* defines the even distribution of the elements of a scene [5]. Equally, *continuity* asks if a sound scene depicts any gaps, if it is uniformly displayed or if it concentrated in "strange places" [19]. Additionally, Toole [19] also uses term *abnormal effects* to denote spatial sensations that do not occur in common experiences. Lindau et al. [7] mention *ghost sources* as an abnormal effect where "[...] sound sources are perceived as ambivalent in direction." [7]. Perhaps the relatively ill-defined sensation of *texture* should also be mentioned at this point [16, 23, 31], which bridges contextually with texture in the context of timbre [14].

A series of terms surrounding diffuseness can be found in the literature as well. *Diffusion* itself is cited to describe the degree to which sound arrives from all directions equally [16]. Likewise, *reverberance* has been used to denote the temporal slurring of an audio event, mainly due to late reflections [19] Conversely, however, the terms *definition* and *clarity* are defined as the degree to which sound can be perceived as being apart from one another and sometimes stands in opposition to *blur* [7, 16, 19].

Finally, there is another wealth of terms that can be found in the artistic research practices. These are often descriptive and originate as conceptualizations of the artists intentions. However, in how far they can be considered valid descriptors for psychoacoustic experiments is even harder to determine. For example, escriptions such as the organ bass pedal note *swirling* through the cathedral [32], or comparing sound spatialization to *sculpture* [22] seem more poetic than measurable. Smalley [23] provides an extensive framework to think of auditory spaces as *perspective* or *zonal* spaces, temporally collapsed *holistic* space, *source-bonded* spaces that transform through alluded context, *mediated* spaces and *gestural* spaces, all of which somewhat touch on envelopment to some capacity.

1.2 Spatial Sound Synthesis

Decorrelation is a common effect that is associated with many aspects of sound envelopment [2], including concert hall envelopment [18]. If implemented as a process, it transforms one audio signal into multiples that appear different from each other, but sound the same [33]. Leveraging decorrelation effects in multichannel loudspeaker arrays is a technique used extensively to artificially produce sound sources with an apparent extent [34]. Using allpass filters with a random phase response is also a common way to achieve artificial reverberation effects. Moreover, Kendall [33] reports that decorrelation increases diffuseness and stabilizes the auditory image.

Spatial sound synthesis methods all rely on achieving some type of decorrelation. These synthesis approaches are characterized by interweaving the creation of sound spectra with the act of sound spatialisation. There are two classes of spatial sound synthesis methods: those that implement a spectral synthesis method first and apply a spatialization approach after, and others that achieve sound spectrum synthesis through the act of spatialization directly. Fully enveloping sounds can be created in both cases.

Spectral spatialization splits the frequency content of a source into several frequency bands and assigns a speaker or spatial position to each band respectively [25]. Without any further processing it should be noted that, perceptually, one does not perceive the individual frequency bands as separate from each other. Instead, spectral integration supersedes spatial segregation [35], meaning that individual frequency components distributed in space fuse together and the sound as a whole is perceived with an extended, enveloping quality.

Spatial granular synthesis subdivides the source sound into grains in the temporal domain [10, 26], similar to classic granular synthesis. However, each grain is assigned a loudspeaker or position in space, similar to the frequency bins of spectral spatialization. The assignment can be done on a per-grain basis or each grain can be sent to predefined audio streams, which can act as the virtual sound sources of an abstraction layer.

While both spectral spatialization and spatial granular synthesis fall into the first class of spatial sound synthesis methods, Spatial Matrix (SM) synthesis [11] and Spatial Modulation synthesis [27] fall into the latter. SM synthesis achieves sound modulation effects comparable to AM and FM synthesis through high velocity sound rotations at audio rate [11]. A basic implementation of SM synthesis in 2D can be done by implementing a rotation matrix operation on an Ambisonic stream at audio rate. Similarly, Spatial Modulation synthesis exploits spatial trajectories at audio rate, but the analysis of which is focused on changes in distance of the source to the listener, exploiting Doppler shifts that yield effects also comparable to FM synthesis [27].

2 Methods

The experiment employs a relative scoring method, asking each participant to score a number of spatialization approaches, henceforth shortened to *algorithms*, according to each of the selected perceptual dimensions. The approach is inspired by the ITU recommendation BS.1534 (MUSHRA) [36]. However, this investigation does not investigate the fidelity of a particular spatial reproduction technique. Thus, no proper reference against which to compare each condition against exists and the participant was asked to compare the conditions to each other, similar to Kuusinen and Lokki [29]. Hypothesized edge-case stimuli were hidden among the multiple options, representing potential positive or negative anchors depending on whichever perceptual dimension is measured. These edge-cases were also used to determine if a listener could be considered an expert. See section 2.3 for details.

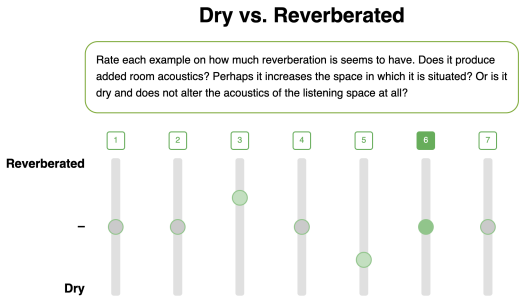


Fig. 1: One example page of the experiment's UI.

There were a total of 3 source sounds spatialized using 7 algorithms over a horizontal ring of 8 channels, with 2 extra channels at 0° and 180°. This created 21 conditions. Effects related to height, such as engulfment [12, 28] or distance [27] could not be considered in this experiment. The conditions were grouped by source sound and perceptual dimension, covering all conditions for a single source before continuing to the next. A trial consisted of presenting all conditions for one dimension and source on a page for evaluations, as shown in Fig. 1. The participants scored all 6 perceptual dimensions before advancing to the next source sound. The participants were aided by providing a brief description for each dimension once more, as well as textural indicators what the markers on the sliders' scale represents. However, the meaning behind each slider remained hidden, requiring the participant to score each algorithm purely on their audition. The order of the sources, as well as the order of the dimensions within each source group and the order of sliders in each trial were randomized for each participant.

A touchscreen was used to allow direct interaction with the UI. On each page, the UI depicted 7 sliders on a scale from 0 to 100, representing each algorithm (see Fig. 1). The participants were not presented with any numerical value while scoring. The source sound looped constantly, allowing the participants to listen to each condition at their own leisure. The UI allowed the participant to mute and unmute each condition. Muting the currently playing condition allowed the participant to rest their ears, but unmuting a new condition would always mute any currently playing one. Only the slider of the currently playing condition was active to avoid any accidental changes of other conditions. The participant could only advance to the next page if all 7 sliders were moved. All sliders started at the 50 score mark and to score a condition at 50 requires the participant to move the slider twice to unlock the 'continue' button, of which they were informed about. A page indicating that a new sound source would now start playing was shown before each source sound change.

Before the start of the experiment, the participants were informed on the nature of the study and the details of each perceptual dimension. They were instructed to abstract from the spectral content and focus on the spatial appearance. However, absolutely no details on how the audio would be spatialized was provided. The participants were able to discuss the dimensions to identify and avoid any possible misconceptions. Furthermore, they completed one test run using a 4^{th} source sound, after which any last doubts on the comprehension of the perceptual dimensions could be discussed. The participants were able to move their head freely, which has been corroborated in the literature [37] and was encouraged beforehand. The multichannel audio was produced using the audio programming environment Max and the UI was implemented in HTML/Javascript using Websockets as the communication layer.

2.1 Participants

A total of 18 participants were selected for this study between the ages 24 to 49 years ($\mu = 40, \sigma = 6.8$), 5 of which identified as female, 12 as male and one reported as other/undisclosed. All were right handed and had self-reported normal hearing. All participants volunteered their time and were not paid to undertake the experiment.

The participants were all considered expert listeners and had prior experience with spatial audio. They were either audio professionals or practicing musicians in the electroacoustic music scene. In a brief questionnaire before the experiment, the participants reported having had between 3 to 21 years of experience with 3D audio ($\mu = 10, \sigma = 5.7$), either in a research, engineering or in an artistic context. Furthermore, they reported 0 to 25 years of experience with audio both in a professional context ($\mu = 16, \sigma = 8.1$) and with electroacoustic music ($\mu = 12, \sigma = 7.4$). The few participants with 0 years of experience, either professionally or with electroacoustic music, had 12 to 20 years of experience in the other category respectively.

2.2 Acoustic Environment

Participants were seated in the middle of a 3D audio post-production room with a size of 7.06m by 5.13m by 3.16m. The room is acoustically treated with with absorption panels and diffusers, and is acoustically isolated from the outside world. The reverberation time is relatively short, with an RT_{60} of 0.398s at 125Hz to 0.253s at 8kHz. The average RT_{60} between 500Hz and 1000Hz is around 0.293s.

8 Genelec 8040 loudspeakers were placed in an equidistant, horizontal ring, rotated so that the first two channels reside at $\pm 22.5^\circ$. Two additional 8040s were placed at 0° and 180° for the front and back conditions. More, unused loudspeakers were visibly present, including on the ceiling. The loudspeakers' height was adjusted to 1.20m, corresponding to an average seated person. The chair was further height adjusted before beginning each experiment. The loudspeakers were calibrated to 75dBSPL using pink noise at -20dBFS at the center listening position.

2.3 Spatialization Approaches (Algorithms)

Monophonic front and back. In two cases, audio was played over a single loudspeaker in front and behind the listener using the two extra loudspeakers in these positions. The *front* condition is supposed to function as a hidden anchor, and act as a control to see if the participant understood the assignment correctly. For the *back* condition we would expect that increased localization blur should result in a slightly heightened sense of envelopment when compared to the front.

Copy. In the *copy* case the source sound was naively played through all 8 loudspeakers equally, creating a correlated sum around the participant. Being a simple

approach to spatialization, it should serve as a another baseline to compare the other methods against.

Decorrelated. For the *decorrelated* case, 8 random phase FIR all-pass filters of length 256 were created. Each one was applied to the output intended for each loudspeaker in the array respectively. The range of the random phase response was $\pm\pi/4$ with a random offset applied per filter. Due to the random generation, many more filters were originally generated and then handpicked based on the criteria that the coloration was minimal, while the perceptual separation of audio signals was maintained.

Spectral Spatialization. The signal was split into 40 bands according to the equivalent rectangular bandwidth starting at a center frequency of 34.197Hz, spanning the entire human hearing range. The bands were uniformly distributed across the ring of 8 loudspeakers at random. This resulted in 5 frequency bands in each loudspeaker. A certain spatial skew may arise, depending on the audio material and band distribution. To counteract this effect new random assignments were created regularly and cross-faded continuously. In order to avoid this cross-fade from becoming noticeable, it was determined during the pilot study that a cross-fade speed of 1 minute was adequate. While slow, participants would usually spend more time evaluating each perceptual dimension, which avoided potentially "unfavorable" frequency band distributions in each page.

Granular Spatialization. For the sake of repeatability, a simple implementation was opted for, dividing the input audio into two constantly overlapping granules using sinusoidal windows. Each granule is then assigned one of 8 loudspeakers in the ring at random. Sine windows were chosen due to summing them maintains the signal's loudness, avoiding any loudness fluctuations. A granule length of 20ms was selected, meaning that a new loudspeaker was chosen for each granule every 10ms in a staggered fashion. This agrees with the literature [10], in that it should still be fast enough to maintain a heightened sense of envelopment.

SM Synthesis. A version of 2D rotational *SM* synthesis was implemented with an audio rate z-axis rotation, whose Ambisonic output was decoded to the 8 channel loudspeaker array. Due to the rotational frequencies of SM synthesis being in the audible range, side-bands are produced in the first order spherical harmonics [11]. Because of this, all audio played through the other algorithms is first amplitude modulated to match the spectral modifications done by SM synthesis. SM synthesis, however, received the unmodulated input audio, as it applies the equivalent modulation through its spatialization itself. To achieve FM synthesis equivalent spectral results for the synthesized scale, the movement needs to be changed from constantly rotating to oscillating around the circle. More details on the source sound generation can be found in section 2.5.

2.4 Perceptual Dimensions

The perceptual dimensions were selected with the summary in section 1.1 as well as the targeted algorithms in mind. As discussed in section 1, the terminology describing attributes of envelopment is insufficient. Nevertheless, the study also needed to be constrained to not cause fatigue with the participants. In the end 6 perceptual dimensions were chosen within which the participants had to evaluate each algorithm and source.

Envelopment. Despite every other measure discussed here relating back to envelopment, the feeling of being surrounded itself is still a valid measure. The scale was depicted as ranging from point source (0) to enveloping (100). It was expected that the monophonic conditions should score lowest on this measure, with the back condition being slightly more enveloping.

Cohesion. In order to see how continuous, or smooth, each algorithm would be, the term cohesion was used to see if the result would appear as a cohesive whole or a disjointed agglomeration. The scale was depicted as ranging from cohesive (0) to disjointed (100). While the monophonic conditions should also score low, the granular spatializaiton condition was expected to feel particularly disjointed.

Distance. In this dimension, the participants were asked to aid themselves in the visible distance to the loudspeaker ring around them and score any percept ± 50 accordingly. The scale was depicted as ranging from close (0) to distant (100). It was hypothesized that in this case the decorrelated algorithm should depict an increase in distance.

Reverberation. Even though no algorithm explicitly adds reverberation, the decorrelation condition could be considered a rudimentary form of reverberation. The participants were told that an increased sense of reverberation from one condition to the other would relate to the impression of the room becoming bigger. If the

condition felt like it added no reverberation apart from the real, dry studio acoustics present, they should mark it at 0. The scale was depicted as ranging from dry (0) to reverberant (100).

Stability. Considering that a few algorithms exhibit quite dynamic source distributions, stability should measure the degree of erratic behavior exhibited by granular spatialization or SM synthesis. The scale was depicted as ranging from stable (0) to changing (100). Naturally, the mono conditions should score low on this measure.

Confusion. Similar to skew [5] or the abnormal effects described by Toole [19], this measure intends to grade the experimental nature of the algorithms. Participants were simply asked to grade how confused they felt about what might be happening in each condition. Explanations included potential skewing or spatial discontinuities. The scale was depicted as ranging from familiar (0) to experimental (100). This measure was chosen well aware of the fact that the participants had extensive prior experience with 3D audio to make a qualified judgment.

2.5 Source Sounds

The choice of source material can profoundly influence the results when evaluating spatialization approaches [5]. A total of 3 source samples were chosen with the intention to both span a wide enough scope but also keep the experiment within reasonable bounds. Those were pink noise, an FM synthesized C-major scale starting at middle C4 and a sample of spoken words in the Cherokee language. While the noise and voice sounds are spectrally richer compared to the synthesized scale, the voice should also represent a real-world example without falling into a speech perception modality. Additionally, a forth cello sample was used to provide the participants with a practice run before they would start the experiment. Given the calibrated 8+2 speaker array, all output combinations between source sound and algorithm were loudness matched.

Constant rotation SM synthesis, produces spectral side bands akin to AM synthesis [11]. All source sounds played through SM synthesis would be AM modulated by the angular speed of rotation. To avoid these spectral differences, all source sounds played in the other conditions are first AM or FM modulated to match the SM synthesis spectrum.

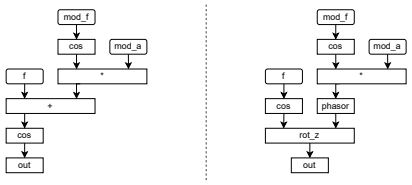


Fig. 2: The FM synthesis implementation used in this study (left) against the equivalent SM synthesis implementation (right).

In the case of the synthesized scale, the signal was modulated using the FM technique to achieve a richer spectrum. Because SM synthesis is capable of producing the same spectrum as FM synthesis by rotating the sound source alone, the generation of the source sound had to be split and be treated as a separate case for SM Synthesis. Generating FM modulated signals with SM synthesis requires the panning position signal to oscillate instead of linearly increase. The two FM implementations are depicted in Fig. 2. In traditional FM the carrier's frequency f is modulated by an amplified sine tone of frequency $\text{mod}_f = 200$ and amplitude $\text{mod}_a = 200$. The identical spectrum is achieved with SM synthesis rotating a source with frequency f around the z-axis, alternating the rotation direction from clockwise to counter-clockwise continuously. In this case, $\text{mod}_a = 200$ is equivalent to the rotation speed, while $\text{mod}_f = 200$ is the frequency with which the source is oscillating in direction.

The voice sample was taken from Max, called `cherokee.aif`. The origin of the sample could be traced back to being part of an Audiosparx sampling library named "Catch 22 SFX" by UK composer Bobby Cole.[1]

2.6 Data Analysis

The self reported scores were sent via a Websocket connection and recorded to a text-file in Max. The compilation and preparation of the data, as well as the descriptive analysis was done in Python using the Numpy and Matplotlib packages. We conducted a repeated-measures ANOVA with bonferroni correction using IBM SPSS Statistics to examine the effects of source type and algorithm on the 6 perceptual dimensions-related dependent measures across 18 participants.

[1] `https://www.audiosparx.com/sa/summary/play.cfm/crumb.1/crumc.0/sound_iid.563341`

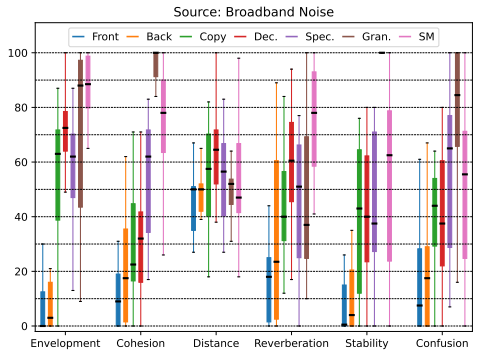


Fig. 3: Boxplots showing quartiles of the relative scores per perceptual dimension across all participants listening to noise.

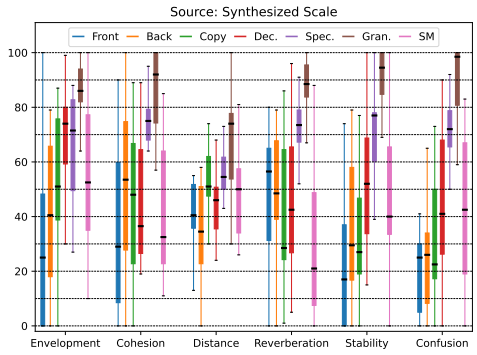


Fig. 4: Boxplots showing quartiles of the relative scores per perceptual dimension across all participants listening to the synthesized scale.

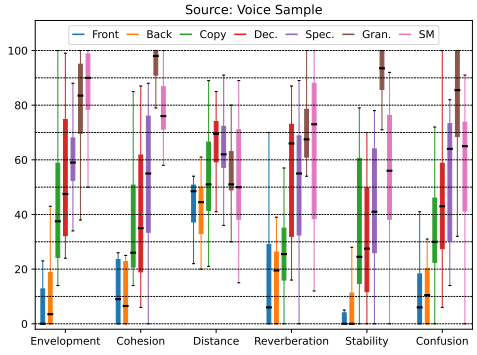


Fig. 5: Boxplots showing quartiles of the relative scores per perceptual dimension across all participants listening to the voice sample.

3 Results and Discussion

The descriptive results are shown in Figs. 3–5. Each plot shows the results for one source across all perceptual dimensions. The 7 boxplots for each algorithm are grouped by perceptual dimension. Within each dimension and for each algorithm the median score is shown inside each box, as well as the 1st to 3rd quartile range (25th to 75th percentile) by the box itself. The whiskers indicate 1.5 times the inter-quartile range.

A series of repeated measures ANOVAs were conducted to evaluate the effects of source sound (3 levels) and spatialization algorithm (7 levels) on the 6 perceptual dimensions using a 3x7 within-subjects design. The Greenhouse-Geisser correction was applied in those cases where Mauchly's sphericity test turned out to be significant. Across dimensions, participants' ratings were significantly influenced by the spatialization algorithm, and in several cases by the source and their interaction.

Envelopment. Significant main effects were found for source ($F(2,34) = 5.77$, $p = .007$) and algorithm ($F(6,102) = 42.38$, $p < .001$), while the interaction was not significant. Post-hoc analysis revealed that granular achieved the highest envelopment ratings, followed by SM and decorrelation, while front received the lowest. This indicates that certain algorithms can enhance the sensation of envelopment consistently across content types.

Cohesion. A significant main effect of the algorithm was observed ($F(6,102) = 28.83$, $p < .001$), with granular yielding the highest cohesion ratings and front the lowest. The source x algorithm interaction was also significant ($F(12,204) = 6.15$, $p < .001$), while the main effect of source was not. These results indicate that cohesion judgments depended on the algorithm used, and were modulated by the specific source sound presented.

Distance. Significant main effects were found for both source ($F(2,16) = 11.02$, $p = .001$) and algorithm ($F(6,12) = 14.35$, $p < .001$), with a significant interaction ($F(12,6) = 3.56$, $p = .049$). SM received the highest distance ratings, followed by granular. The noise source was rated significantly higher in distance than the synthesized scale and voice. These findings indicate that both the type of content and the spatial rendering strategy influenced distance perception.

Reverberation. There were significant main effects of source ($F(2,34) = 6.02$, $p = .006$) and algorithm

($F(6,102) = 19.94$, $p < .001$), as well as a significant interaction ($F(12,204) = 10.56$, $p = .005$). Granular received the highest reverberation ratings and front the lowest, with algorithmic effects differing across stimuli.

Stability. The main effects of source ($F(2,16) = 6.97$, $p = .003$) and algorithm ($F(6,12) = 51.86$, $p < .001$) were significant, as was the source x algorithm interaction ($F(12,6) = 3.37$, $p = .030$). Granular again led to the highest ratings, with SM and spectral also scoring high. Front and back both scored low once more. These results show that the perception of stability was shaped by both the algorithm and the characteristics of the source.

Confusion. We found a strong main effect of algorithm ($F(6,12) = 36.81$, $p < .001$) and a source x algorithm interaction ($F(12,6) = 2.14$, $p = .016$). The main effect of source sound was not significant. Granular produced the highest confusion ratings, followed by SM and spectral, while front and back yielded the lowest. This suggests that perceived confusion was highly sensitive to the algorithm and context-dependent.

Across all perceptual dimensions, the spatialization algorithm had a significant effect, with granular consistently achieving the highest ratings for envelopment, cohesion, reverberation, stability, and confusion. SM also performed strongly, especially for distance. Source type significantly affected ratings in envelopment, distance, stability and reverberation, and interactions between source and algorithm were found in cohesion, distance, reverberation, stability and confusion. These findings underscore the critical role of algorithmic spatialization in shaping the perceived quality and structure of spatial audio experiences.

The results between the noise and voice source are consistent. Participants successfully identified the monophonic frontal anchor across all perceptual dimensions for the noise and voice source. The synthesized scale, however, reveals the importance of testing against several stimuli types. In this case, the stimulus had a limited bandwidth, which could explain the increased difficulty to correctly identify the anchors and evaluate against the perceptual dimensions.

While it might seem, judging by Figs. 3-5, that the back position scores slightly higher compared to the front on many perceptual dimensions and source sounds, the results are not significant enough to draw any conclusions (see Fig. 6). The copy and decorrelation algorithms scored similarly across many perceptual dimensions, with copy mostly being slightly below decorrelation, as expected. However, we could not observe a general benefit in stability over the copy algorithm as previously reported for decorrelated signals [33]. Decorrelation seemed to increase reverberation for the noise and voice sources, but not for the synthesized scale. It also seemed to increase the sense of distance only slightly compared to copy and surprisingly seemed to be rated closer than copy using the synthesized scale. With copy being correlated from all directions, it is also surprising that participants generally did not rate it closer.

The granular algorithm acted as an anchor in opposition to the front position in many perceptual dimensions. The spectral algorithm exhibited more varied scoring, being more comparable to copy and decorrelation for envelopment and stability for the noise and voice sources, but closer to granular and SM in terms of confusion and cohesion. The SM algorithm shows dramatic differences between the synthesized scale and the noise or voice sources. While it generally compares to granular in latter cases, it seems to set itself apart perceptually in the former.

Fig. 6 shows the pairwise comparison results, marking the significant effects across all perceptual dimensions and source sounds. For one, it demonstrates that there are no significant differences between the front and back positions. The copy and decorrelation are also nearly identical, short of reverberation. Of primary interest though the fact, that is the 3 spatial sound synthesis algorithms demonstrate sufficient differences between each other, which may indicate that within the selected 6 perceptual dimensional perceptual space each approach is indeed capable of producing unique enveloping spatial impressions.

4 Summary

This study presented a comparison of 7 approaches to sound spatialization over 6 perceptual dimensions using 3 source sounds over a ring of 8 loudspeakers, plus 2 extra loudspeakers in the front and back position. While 2 spatialization approaches were monophonic from the front and back, 3 spatial sound algorithms were chosen, as well as simply playing the source through all 8 loudspeakers both correlated and decorrelated. 18 participants rated each condition on a 0–100 scale. The results suggest that the choice of spatialization algorithm

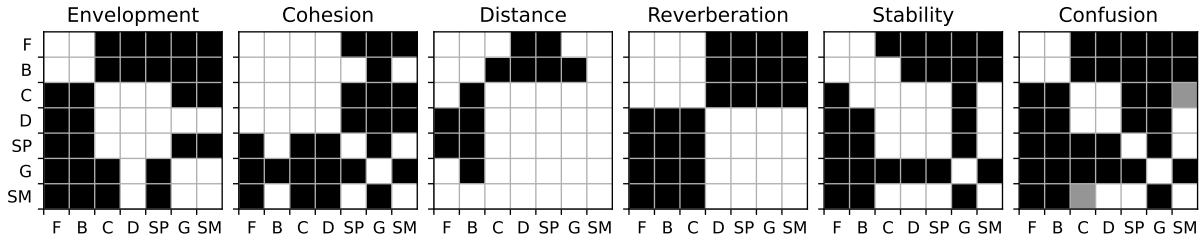


Fig. 6: Pairwise comparison significance matrices. Black: $p <= .05$; grey: $.05 < p <= 0.1$; and white: $p > 0.1$.

used has significant effects on the perception of envelopment, cohesion, distance, reverberation, stability and confusion. There were also some significant effects for the choice of source sound. Stark differences could be observed between the synthesized scale source vs the others. A pairwise comparison indicates that the algorithms may indeed be able to produce unique spatial sensation each. However, more work will be required to confirm these findings.

The study was not without shortcomings, however. The correct choice of perceptual dimensions is complicated, as demonstrated in the literature review. While the results were satisfactory and significant, they do not fully prove the adequacy of the choice entirely. Some perceptual dimensions, such as confusion, were still slightly confusing in itself. During the post-experiment discussion participants revealed that they listened to both the experimental nature as well as unwanted artifacts, suggesting a further break down into smaller perceptual descriptors might be necessary. Also, despite best efforts, there is suspicion that some participants mixed up stability with robustness. Furthermore, having different perceptual dimensions be represented differently in the UI, such as distance, was not ideal and required extra care to avoid confusion.

Interestingly, SM synthesis was used with both constant rotations and its oscillating version. The results indicate some differences between the produced stimuli in each case. Future work will be necessary to unravel the effect of source sound from the version of SM synthesis used to confirm if expert listeners are indeed capable of distinguishing between both versions spatially. However, it also stands that the synthesized scale source sound does not support several hypotheses in particular. This may be due to the sparse spectrum generated in this case.

In general, further standardization is needed to imbue more confidence, as many results are implementation specific, from the length of the decorrelation filters to, especially, the many ways spectral, granular or SM synthesis can be implemented. While this explorative study intends to gain a broad oversight over the many spatializations algorithms and envelopment dimensions, it's results can only reflect the specific implementations of the selected spatialization methods, identified during pilot studies. Thus, the results presented here show that these specific approaches to spatialization have significant effects on the perception of envelopment. While not invalidating the results of this study, other parameter choices may indeed have different effects on the perception of envelopment. Further investigations into the finer details using each spatialization approach must therefore be part of future work.

References

[1] Morimoto, M., Iida, K., and Furue, Y., "Relation between auditory source width in various sound fields and degree of interaural cross-correlation," *Applied Acoustics*, 38(2–4), pp. 291–301, 1993.

[2] Blauert, J., *Spatial hearing: the psychophysics of human sound localization*, MIT press, Cambridge, MA, USA, 1997.

[3] Lokki, T. and Pätynen, J., "Auditory spatial impression in concert halls," in *The Technology of Binaural Understanding*, pp. 173–202, Springer, 2020.

[4] Lehnert, H., "Auditory spatial impression," in *12th International Conference: The Perception of Reproduced Sound*, Copenhagen, Denmark, 1993.

[5] Rumsey, F., "Spatial quality evaluation for reproduced sound: Terminology, meaning, and a scene-based paradigm," *JAES*, 50(9), pp. 651–666, 2002.

[6] Berg, J., "The contrasting and conflicting definitions of envelopment," in *126th Audio Engineering Society Convention*, 2009.

[7] Lindau, A., Erbes, V., Lepa, S., Maempel, H.-J., Brinkman, F., and Weinzierl, S., "A spatial audio quality inventory (SAQI)," *Acta Acustica united with Acustica*, 100(5), pp. 984–994, 2014.

[8] Weinzierl, S., Lepa, S., and Ackermann, D., "A measuring instrument for the auditory perception of rooms: The Room Acoustical Quality Inventory (RAQI)," *The Journal of the Acoustical Society of America*, 144(3), pp. 1245–1257, 2018.

[9] Kendall, G. S., "Spatial perception and cognition in multichannel audio for electroacoustic music," *Organised Sound*, 15(3), pp. 228–238, 2010.

[10] Riedel, S., Frank, M., and Zotter, F., "Perceptual evaluation of listener envelopment using spatial granular synthesis," *arXiv preprint arXiv:2301.10210*, 2023.

[11] Schmele, T. and Garriga, A., "Spatial Matrix Synthesis," *Journal of the Audio Engineering Society*, 72(4), pp. 193–210, 2024.

[12] Lynch, H. and Sazdov, R., "A perceptual investigation into spatialization techniques used in multichannel electroacoustic music for envelopment and engulfment," *Computer Music Journal*, 41(1), pp. 13–33, 2017.

[13] Spors, S., Wierstorf, H., Raake, A., Melchior, F., Frank, M., and Zotter, F., "Spatial sound with loudspeakers and its perception: A review of the current state," *Proceedings of the IEEE*, 101(9), pp. 1920–1938, 2013.

[14] Erickson, R., *Sound structure in music*, University of California Press, Berkley, CA, USA, 1975.

[15] Beranek, L. L., "Music, Acoustics, and Architecture," *Bulletin of the American Academy of Arts and Sciences*, 45(8), pp. 25–46, 1962.

[16] Hawkes, R. and Douglas, H., "Subjective acoustic experience in concert auditoria," *Acta Acustica united with Acustica*, 24, pp. 235–250, 1971.

[17] Bradley, J. S. and Soulodre, G. A., "The influence of late arriving energy on spatial impression," *Journal of the Acoustical Society of America*, 97(4), pp. 2263–2271, 1995.

[18] Cabrera, D., Nguyen, A., and Choi, Y.-J., "Auditory Versus Visual Spatial Impression: A Study of Two Auditoria." in *10th Internationa Conference on Auditory Display*, Sydney, Australia, 2004.

[19] Toole, F. E., "Subjective measurements of loudspeaker sound quality and listener performance," *Journal of the Audio Engineering Society*, 33(1/2), pp. 2–32, 1985.

[20] Potard, G. and Burnett, I., "A study on sound source apparent shape and wideness," in *International Conference on Auditory Display*, Boston, MA, USA, 2003.

[21] Emmerson, S. and Smalley, D., "Electro-acoustic music," in *Oxford Music Online*, Oxford University Press, 2001.

[22] Harrison, J., "Sound, space, sculpture: some thoughts on the 'what','how'and 'why'of sound diffusion," *Organised sound*, 3(2), pp. 117–127, 1998.

[23] Smalley, D., "Space-form and the acousmatic image," *Organised sound*, 12(1), pp. 35–58, 2007.

[24] Normandeau, R., "Timbre Spatialisation: The medium is the space," *Organised Sound*, 14(3), pp. 277–285, 2009.

[25] Kim-Boyle, D., "Spectral spatialization – an Overview," in *International Computer Music Conference*, Belfast, Northern Ireland, 2008.

[26] Wilson, S., "Spatial swarm granulation," in *International Computer Music Conference*, Belfast, Northern Ireland, 2008.

[27] McGee, R., "Spatial modulation synthesis," in *International Computer Music Conference*, pp. 246—-249, Denton, TX, USA, 2015.

[28] Sazdov, R., Paine, G., and Stevens, K., "Perceptual Investigation into envelopment, spatial clarity, and engulfment in reproduced multi-channel audio," in *31st International Conference of the AES*, London, UK, 2007.

[29] Kuusinen, A. and Lokki, T., "Investigation of auditory distance perception and preferences in concert halls by using virtual acoustics," *The Journal of the Acoustical Society of America*, 138(5), pp. 3148–3159, 2015.

[30] Barron, M., "The subjective effects of first reflections in concert halls—the need for lateral reflections," *Journal of Sound and Vibration*, 15(4), pp. 475–494, 1971.

[31] Hagan, K. L., "Textural composition: Aesthetics, techniques, and spatialization for high-density loudspeaker arrays," *Computer Music Journal*, 41(1), pp. 34–45, 2017.

[32] Griesinger, D., "General overview of spatial impression, envelopment, localization, and externalization," in *15th International Audio Engineering Society Conference*, pp. 136–149, Copenhagen, Denmark, 1998.

[33] Kendall, G. S., "The decorrelation of audio signals and its impact on spatial imagery," *Computer Music Journal*, 19(4), pp. 71–87, 1995.

[34] Potard, G. and Burnett, I., "Decorrelation techniques for the rendering of apparent sound source width in 3D audio displays," in *International Conference on Digital Audio Effects*, pp. 280–284, Naples, Italy, 2004.

[35] Bregman, A. S., *Auditory scene analysis: The perceptual organization of sound*, MIT press, Cambridge, MA, USA, 1990.

[36] ITU-R Recommendation BS.1534, "Method for the subjective assessment of intermediate quality levels of coding systems," 2015.

[37] Martens, W. L. and Han, Y., "Discrimination of Auditory Spatial Diffuseness Facilitated by Head Rolling While Listening to With-Height versus Without-Height Multichannel Loudspeaker Reproduction," in *AES International Conference on Spatial Reproduction-Aesthetics and Science*, Tokyo, Japan, 2018.