Chapter

The Influences of Hearing and Vision on Egocentric Distance and Room Size Perception under Rich-Cue Conditions

Hans-Joachim Maempel and Michael Horn

Abstract

Artistic renditions are mediated by the performance rooms in which they are staged. The perceived egocentric distance to the artists and the perceived room size are relevant features in this regard. The influences of both the presence and the properties of acoustic and visual environments on these features were investigated. Recordings of music and a speech performance were integrated into direct renderings of six rooms by applying dynamic binaural synthesis and chroma-key compositing. By the use of a linearized extraaural headset and a semi-panoramic stereoscopic projection, the auralized, visualized, and auralized-visualized spatial scenes were presented to test participants who were asked to estimate the egocentric distance and the room size. The mean estimates differed between the acoustic and the visual as well as between the acoustic-visual and the combined single-domain conditions. Geometric estimations in performance rooms relied upon nine-tenths on the visual, and one-tenth on the acoustic properties of the virtualized spatial scenes, but negligibly on their interaction. Structural and material properties of rooms may also influence auditory-visual distance perception.

Keywords: auditory-visual perception, virtual reality, egocentric distance, room size, performance room, concert hall, music, speech

1. Introduction

1.1 Desideratum

The multimodal perception, integration and mental reconstruction of the physical world provide us, amongst other things, with various modality-specific and modality-unspecific features such as colors, timbres, smells, vibrations, locations, dimensions, materials, and aesthetic impressions, which are or can be related to perceived objects and environments. A fundamental issue is the extent to which such features rely on the different modalities and their cooperation. The present study examined and experimentally dissociated the important modalities of hearing and vision by separately providing and manipulating the respectively perceivable information about the physical world, i.e., auralized and visualized spatial scenes. In everyday life, both the egocentric distance to visible sound sources and the size of a

surrounding room are important perceptual features, since they contribute to spatial notion and orientation. They are also relevant about artistic renditions and performance rooms, as they relate, for instance, to the concept of auditory intimacy, an important aspect of the quality of concert halls [1–3]. Accordingly, both the perceived egocentric distance and the perceived room size were investigated, primarily in the context of artistic renditions.

1.2 State of the art

The interaction between hearing and vision occurs in the perception of various features, pertaining for example to intensity [4], localization [5–7], motion [8–10], event time [11, 12], synchrony [13], perceptual phonetics [14], quality rating [15], and room perception [16–18]. Regarding auditory-visual localization and spatial perception, research has focused mainly on horizontal directional localization to date, followed by distance localization, while room size perception has rarely been investigated. Two superior research objectives may be identified in the literature: One objective is the description of human perceptual performance and its dependence on physical cues. Within this context, distance perception was mainly investigated about its *accuracy*, and specifically via the experimental variation of the cues about the equivalent physical distance. The consideration of interfering factors such as the completeness and the integrity of the cues may be subsumed under this objective, too. Another objective is the modeling of internal processes of multisensory integration, which are closely related to the binding problem. The binding problem asks, how different sensory information is identified as belonging to the same event, object or stream, and thus is unified. According to Treisman there are "at least seven different types of binding": property, part, range, hierarchical, conditional, temporal, and location binding ([19], p. 171).

Experimental stimuli may be real objects (e.g., humans, loudspeakers, mechanical apparatuses) that have diverse physical properties and may bear meaning. Otherwise, the investigation of detailed internal mechanisms using behavioral experiments often calls for neutral objects or energetic events with a maximally reduced number of properties and without meaning (e.g., lights, noise) [5]. Criteria for the selection of one of these stimulus categories are essentially the options of stimulus manipulation (e.g., real objects will hardly allow for conflicting stimuli) and the relation of internal and external validity. The advancement of virtual reality provided experimenters with extended and promising options for manipulating complex, naturalistic stimuli. Since the virtualization of real environments is known to affect various perceptual and cognitive features [20–23], the impact of virtualization has become another prominent research issue.

The perception of distance and room size in the extrapersonal space depends on particular auditory and visual cues provided by the specific scene. Acoustic distance cues are weighted variably and comprise the sound pressure level and the direct-to-reverberant energy ratio [24–26], spectral attenuation due to air absorption [27], spectral properties due to temporal and directional patterns of reflections of surrounding surfaces [25], as well as spectral alterations due to both near-field conditions and the listener's head and torso. Interaural level and time differences also appear to play a role, namely in connection with orientations and motions of the sound source and the listener [28–30].

In real acoustic environments, perceived egocentric distances are known to be compressed above distances of 2 to 7 m [27, 28, 31–33], hence they are found to be compressed comparably or even more in virtual acoustic environments [32, 34–37]. However, a largely accurate estimation in high-absorbent and an overestimation in low-absorbent virtual environments were also reported [18, 38].

Acoustic room size cues comprise the room-acoustic parameters clarity (C80, C50) [39–41], definition (D50) [41], reverberation time (RT) [39, 42, 43], and likely the characteristics of early reflections [39]. In the medium- and large-sized rooms, the perceived room size was shown to be decreased by a binaural reproduction of the acoustic scene compared to listening in situ [40]. A more recent study found, however, that auralization by dynamic binaural synthesis did not affect the estimation of room size [38].

The estimation of the egocentric distance and the dimensions of visual rooms is based on visual depth cues. Common classifications differentiate between pictorial and non-pictorial, monocular and binocular, as well as visual and oculomotor cues. The cues cover different effective ranges: the personal space (0–2 m), the action space (2–30 m) and/or the vista space (> 30 m) [44]. The non-pictorial depth cues comprise three oculomotor cues: *Convergence* refers to the angle between the eyes' orientation towards the object, *accommodation* to the adaptation of the eye lens' focal length, and *myosis* to the pupillary constriction. *Convergence* is the only binocular oculomotor cue. *Myosis* is effective only within the personal space. Further important non-pictorial visual depth cues are *binocular parallax* (also termed binocular/retinal disparity) referring to differences between the two retinal images due to the permanently different eyes' positions, and *monocular motion (movement) parallax* referring to subsequently different retinal images due to head movements. These cues are effective in both the personal and the action space. Pictorial depth cues are always monocular and based on the extraction of features from the specific images and, where applicable, experiential knowledge. *Linear perspective*, *texture gradient*, *overlapping (occlusion)*, *shadowing/shading*, *retinal image size*, *aerial perspective* and the *height in the visual field* appertain to the most important pictorial depth cues (see [44–46] for an overview).

In real visual environments, distances are normally estimated much more precisely and accurately than in real acoustic environments [47]. Beyond about 3 m distances are increasingly underestimated both under reduced-cue conditions [48] and in virtual visual environments, no matter if head-mounted displays or large screen immersive displays are used [38, 49–55]. However, also largely accurate estimates in virtual visual environments were reported [18, 56]. While the parallax and the observer-to-screen distance [57], as well as stereoscopy, shadows, and reflections [58] were identified to influence the accuracy of distance estimates in virtual visual environments, the restriction of the field of view [59] and the focal length of the camera lens [60] did not take effect. Room size was observed to be overestimated more in a real visual environment than in the correspondent virtual environment [38], as well as underestimated in other virtual visual environments [18].

Turning to acoustic-visual conditions, the experimental combination of acoustic and visual stimuli can be either congruent or divergent regarding positions or other properties. The widely-used variation of the *presence* of congruent stimulus components (acoustic/visual/acoustic-visual) may be referred to as a co-presence paradigm. A divergent combination independently varies the acoustic and visual *properties* of an acoustic-visual stimulus and is commonly referred to as a conflicting stimulus paradigm.

Under congruent conditions, as experienced in real life, distance estimation is normally highly accurate. Using virtual sound sources and photographs, the additional availability of visual distance information was demonstrated to improve the linearity of the relationship between the physical and the perceptual distance, and to reduce both the within- and the between-subjects variance of the distance judgments [61]. However, virtual acoustic-visual environments may, like virtual visual environments, be subject to compressed distance perception [32], regardless of the application of verbal estimation or perceptually directed action as a measurement

protocol [36, 37]. A perceptual comparison between mixed and virtual reality [62] showed that the virtualization of the visual environment increased "aurally perceived" distance and room size estimates (p. 4). The perceived room width was found to be underestimated under the visual, overestimated under the acoustic, and well-estimated under the acoustic-visual conditions [17]. Findings on the accuracy of room size perception are in the same way inconsistent for acoustic-visual environments, as they are for visual environments (see above) [18, 38].

Experiments applying the conflicting stimulus paradigm are normally both more challenging and more instructive [36]. Such experiments have revealed that the localization of an auditory-visual object is largely determined by its visual position, which becomes particularly obvious when compared to the localization of an auditory object. This phenomenon was investigated relatively early [5], and in the case of a lateral or directional offset in the horizontal plane, it was initially referred to as the *ventriloquism effect* ([6], pp. 360-2, [63, 64]). This term has been used in a more abstract sense since, refers to both the spatial and the temporal domain, as well as both directional and distance offsets. The respective effects and aftereffects have been extensively studied (see [65] for an overview).

In the case of an egocentric distance offset, the phenomenon was initially termed the *proximity image effect*: In 1968, Gardner reported that in an anechoic room, the perceived distance was fully determined by the distance of the only visible nearby loudspeaker [7]. A modified replication showed that the effect occurred also when the acoustic distance was nearer than the visual distance, and was only slightly weakened by the chosen semi-reverberant conditions [66]. Zahorik did, however, not observe a clear *proximity image effect* in his replication [67]. Rather, auditory-visual perception, allowing also for prior inspection of the potential sound source locations, improved judgment accuracy when compared to auditory perception (see also [33]). The lack of support for a strict visual dominance in auditory-visual distance localization suggested that sensory modalities contribute to localization with scalable weights.

Indeed, it has been demonstrated that both visual and acoustic stimulus displacements cause significant changes in egocentric distance estimates [68], indicating that visual and auditory influences occur at the same time, however, with different weights. Regarding auditory features, Postma and Katz varied both visual viewpoints and auralizations in a virtual theater, while asking experienced participants for ratings upon distance and room acoustic attributes [69]. Few attributes (including auditory distance) were significantly influenced by the visual contrasts, whereas most attributes were by the acoustic. Interestingly, a deeper data analysis allowed partitioning participants into three groups being mainly susceptible to auditory distance, loudness, and none of the features, respectively, when exposed to different visual conditions. Amongst others, the study points to the principle, that acoustic and visual information weigh normally highest on auditory and visual features, respectively.

In the course of the advancement of a probabilistic view, it was evidenced that the weights adapt to the reliabilities of the sensory estimates in a statistically optimal manner [70]. Maximum Likelihood Estimation (MLE) modeling was shown to apply to different multisensory localization tasks [47, 71–73]. Therefore, acoustic-visual stimuli should generally yield a more precise localization than merely acoustic or visual stimuli [72]. The weights may either be experimentally reduced by adding noise to the stimuli, or in turn, if estimated otherwise, indicate the relative acuity of the stimuli and the reliability of their sensory estimates, respectively. For instance, due to missing or largely reduced interaural level difference and interaural time difference cues, auditory positional information has a lower weight in case of a directional or depth offset in the median plane; in this case, localization is therefore more prone to the influence of visual positional information than in the case of a

lateral offset [9, 74]. It was found that acoustic and visual contributions are not symmetric about frontal distance: Using LEDs and noise bursts, a "visual capture" effect and a respective aftereffect in frontal distance perception was observed, with a relatively greater visual bias for visual stimulus components being closer than the acoustic components ([75], p. 4).

Combining MLE with Bayesian causal inference modeling [76] is based on the idea that increasing temporal or spatial divergences between sensory-specific stimuli make the perceiver's inference of more than one physical event more likely, and that multisensory integration takes place only for stimuli subjectively caused by the same physical event. A recent study demonstrated, however, a higher weight of visual signals in auditory-visual integration of spatial signals than predicted by MLE, which might be due to the participants' uncertainty about a single physical cause [77]. While the result of the causal inference is normally not directly observable, the perceived spatial congruency is: Using stereoscopic projection and wave field synthesis, André and colleagues presented participants with 3D stimuli (a speaking virtual character) containing acoustic-visual angular errors. As expected, a higher level of ambient noise (SNR = 4 dB A) caused a 1.1° shift of the point of subjective equivalence and a steeper slope (−0.077 instead of −0.062 per degree) of the psychometric function. Results were not statistically significant, arguably due to the still too high SNR [78].

Evaluating different variants of probabilistic models through experiments using a virtual acoustic-visual environment and applying a dual-report paradigm, the Bayesian causal inference model with a probability matching strategy was found to explain the auditory-visual perception of distance best [79]. The authors also calculated the sensory weights for visual and auditory distances and found that in windows around the correspondent physical distance, auditory distances were predominantly influenced by visual, while visual distances were slightly influenced by auditory sensory estimates. Visual-auditory weights ranged from 0 to 1, auditory-visual weights from 0 to 0.2. Another study showed a major influence of the acoustic properties of spatial scenes on the collective egocentric distance perception (probably due to a substantially restricted visual rendering), whereas room size perception predominantly relied on the visual properties. The virtual environment was based on the dynamic binaural synthesis, speech and music signals, stereoscopic still photographs of a dodecahedron loudspeaker in four rooms, and a 61″ stereoscopic full HD monitor with shutter glasses [18].

The cited studies applied different data collection methods (e.g., triangulated blind walking, absolute scales, 2AFC), virtualization concepts (no virtualization, direct rendering, numerical modeling), stimulus content types (e.g., speech, noise; LEDs, visible sound sources), visual moves (photographs, videos), stimulus dimensionalities (2D, 3D), and reproduction formats (e.g., monophonic sound, sound field synthesis; head-mounted displays, large immersive screens). Thus, connecting the results in a systematic manner is challenging. Findings on the influences of concrete physical properties on percepts and their parameters have not achieved consistency.

Following a research strategy from the general to the specific, the present study focuses on the influences of the acoustic and visual environments' properties in their totality. To this end, whole rooms and source-receiver configurations were experimentally varied. To make this feasible, a collective instead of an individual testing approach was taken, i.e., identical test conditions were allocated not to different repetitions (as necessary for data collection in the context of probabilistic modeling) but to different participants. To emphasize external validity and step towards "naturalistic environments" ([65], p. 805), two prototypic types of content (music, speech), six physically existing rooms, direct 3D renderings, long and meaningful stimuli, and a perceptually validated virtual environment were applied.

1.3 Research questions and hypotheses

Methodologically, the prominent co-presence paradigm entails two restrictions. Firstly, the comparison between the acoustic or visual and the acoustic-visual condition involves two sources of variation: (a) the change between the stimulus' domains (acoustic vs. visual), and (b) the change between the numbers of stimulus domains (1 vs. 2)—i.e., between two basic modes of perceptual processing. Thus, the co-presence paradigm confounds two factors at the cost of internal validity. Since single-domain (acoustic, visual) stimuli do not require a multimodal trade-off, whereas multi-domain (acoustic-visual) stimuli do, different weights of auditory and visual information depending on the basic mode of perceptual processing are expected [79]. To take account of the sources of variation, two dissociating research questions (RQs) were posed.

As a second restriction, the co-presence paradigm does not cover variations within the multi-domain stimulus mode, though it is prevalent in everyday life. Hence, additional RQs ask for the effects of the *properties* of acoustic and visual environments. The respective hypotheses were tested based on six performance rooms with particular source-receiver arrangements, and of both music and speech performances.

RQ 1: To what extent do the perceptual estimates depend on the stimulus domain (acoustic vs. visual, and thereby of the involved modality) as such?
$H1_0$: $\mu_A = \mu_V$.

RQ 2: To what extent do the perceptual estimates depend on the basic mode of perceptual processing (single vs. multi-domain stimuli)?
$H2_0$: $2 \cdot \mu_{AV} = \mu_A + \mu_V$.

RQ 3: To what extent do the perceptual estimates depend on the complex acoustic properties of the multi-domain stimuli?
$H3_0$: $\mu_{A1V\bullet} = \mu_{A2V\bullet} = \mu_{A3V\bullet} = \mu_{A4V\bullet} = \mu_{A5V\bullet} = \mu_{A6V\bullet}$.

RQ 4: To what extent do the perceptual estimates depend on the complex visual properties of the multi-domain stimuli?
$H4_0$: $\mu_{A\bullet V1} = \mu_{A\bullet V2} = \mu_{A\bullet V3} = \mu_{A\bullet V4} = \mu_{A\bullet V5} = \mu_{A\bullet V6}$.

RQ 5: To what extent do the perceptual estimates depend on the interaction of the complex acoustic and visual properties of the multi-domain stimuli?
$H5_0$: $\mu_{AjVk} = \mu_{AjV\bullet} + \mu_{A\bullet Vk} - \mu_{A\bullet V\bullet}$ with $1 \leq j \leq 6$ and $1 \leq k \leq 6$.

Note that not only distance and room size cues but whole scenes were varied, to infer the effects of the entire physical properties of the performance rooms, and therefore of the sensory modalities as such in the context of these environments. RQs 3–5 were made comparative by asking to which extent acoustic and visual properties, and their interaction, do proportionally account for the estimates. For this purpose, commensurable ranges of the factors had to be ensured (2.3, 2.7).

Dependent variables were the perceived egocentric distance and the perceived room size. Where reasonable, the accuracy of the estimates about the physical distances and sizes was also considered.

2. Method

2.1 Methodological considerations and terminology

Answering RQs 1 to 2 requires the application of the co-presence design paradigm. Auralized, visualized, and auralized-visualized spatial scenes are levels of one factor.

Answering RQs 3 to 5 requires the acoustic and visual properties of the scenes to be independent factors rather than just levels of one factor, i.e., the application of the conflicting stimulus paradigm. To allow for the quantification of the proportional influences of acoustic properties, visual properties, and their interaction on the perceptual features, however, certain methodological criteria have to be met, because light and sound cannot be directly compared due to their different physical nature. In particular, not only spatiotemporal congruency but also "crossmodal correspondences" (involving low-level features) ([80], p. 973) as well as semantic congruency [80] of the acoustic and visual stimuli being based on the same scenes (which therefore 'sound as they look'), and the qualitative and quantitative commensurability of the acoustic and visual factors are all required. To this end, the single-domain (acoustic and visual) stimuli have to be derived from the same set of multi-domain (acoustic-visual) stimuli and must be varied in their entirety, i.e., categorically [81].

These considerations result in the need for preservation of all perceptually relevant physical cues and a direct rendering, which we distinguish from fully numerical or partly numerical (hybrid) simulations. The latter approaches are based on assumptions of the physical validity of parametrized material and geometrical room properties, the imperceptibility of structural resolution limits, and/or the physical validity of the applied models on sound and light propagation, including methods of interpolation. By using the term direct rendering, we indicate that the rendering data corresponding to all supported participants' movements were acquired in situ, i.e., neither calculated from a numerical 3D model nor spatially interpolated (see 2.5.).

With the objective of a clear description of investigated effects, it is indicated to factually and terminologically differentiate between ontological realms (*physical*, *perceptual*), and therein between both physical domains (*acoustic*, *visual*; elsewhere also termed *acoustic* and *optical*) and perceptual modalities (*auditory*, *visual*), as well as between modal specificities (*unimodal*, *supramodal*; also referred to as *modal* and *amodal* [80]) [81].

2.2 Perceptual features

In view of both the context of the study (artistic renditions, performance rooms) and the complex variation of the stimuli (2.1), the collection of values of various features was of interest. Accordingly, a differential was used. A superordinate objective of the research project is a comparison of the features regarding their respective dependencies on the presences and properties of the acoustic and visual stimuli. Hence, the questionnaire consisted of 21 perceptual features, subdivided into four sets: auditory features (e.g., *reverberance*), visual features (e.g., *brightness*), aesthetic and presence-related auditory-visual features (e.g., *pleasantness*, *spatial presence*), and geometric auditory-visual features (*source distance*, *source width*, *room length*, *room width*, *room height*). Following [82, 83], reference objects (quartet/speaker, room) of the visual and the geometric features were specified. The features were operationalized by bipolar rating scales which were displayed on a tablet computer. Data were entered using touch-sensitive, graphically continuous sliders with a numerical resolution of 127 steps. The geometric feature scales specified units [m] and ranged from 0 to 5 m (*source width*), to 25 m (*source distance*, *room height*), to 50 m (*room width*), and to 100 m (*room length*). Interval scaling was assumed. The original test language was German. Both the perceived distance and the perceived room size are supramodal (amodal) features by definition [80, 81]. Since optimal preconditions for crossmodal binding and bisensory integration had been established by ensuring crossmodal correspondences and semantic congruency [80, 84], and since they are constant across the co-presence variation and to a considerable extent constant across the conflicting stimulus variation, auditory-

Measure	*Perceived length* $\hat{L}$	*Perceived width* $\hat{W}$	*Perceived height* $\hat{H}$	*Perceived source distance* $\hat{D}$
Mean	45.833	22.162	14.672	10.431
Standard error of mean	0.905	0.542	0.229	0.185
Cronbach's Alpha	0.926	0.889	0.867	0.850

Table 1.
Comparison of descriptives and internal consistencies of the unidimensional perceptual features. Calculations are based on the total sample (music and speech group, N = *88) and all rooms under the mere visual conditions (**V1–V6**). The conditions were pooled for the calculation of mean and standard error, and treated as separate items for the calculation of Cronbach's alpha.*

visual integration was assumed to be able to occur either automatically or intentionally. Hence, participants were asked to estimate values of unitary features. No problems concerning this task were reported. Because test participants do not maintain linearity when assessing three-dimensional room volume using a single one-dimensional scale [18], they were asked for separate length ($\hat{L}$), width ($\hat{W}$), and height ($\hat{H}$) estimates.

Since the visual stimuli showed only a part of the frontal hemisphere (see 2.5), the participants had to base their assessment of the invisible rear part of the rooms' length on the visible frontal length, the room shape, their position in the room, and their experiential knowledge on the shape and size of performance rooms. Hence, before analyzing the calculated room volume/size estimates, dispersion and reliability measures of the unidimensional perceptual features were inspected (**Table 1**).

Neither the reliability nor the dispersion of the perceived length is conspicuous, since the values for Cronbach's Alpha are throughout high, for the perceived length even excellent, and the error-to-mean ratios are consistent across the perceptual features. By calculating the cube root of the product of the three collected features, the one-dimensional feature *perceived room size* $\hat{S}$ was derived. This report focuses on the *perceived source distance* ($\hat{D}$) and the *perceived room size* ($\hat{S}$).

2.3 Design

Since answering RQs 1 to 2 requires the application of the co-presence paradigm, the factor *Domain* was defined by the levels auralized (**A**), visualized (**V**), and auralized-visualized (**AV**). To raise the external validity of the potential main effects and to allow for the observation of room-specific effects, the second factor *Room* was introduced, comprising six different performance rooms under examination (levels **R1** to **R6**, see **Table 2** for specific labels). Answering RQs 3 to 5 requires the application of the conflicting stimulus paradigm. Thus, the factor *Auralized room* was defined by the acoustic stimulus components of the six rooms (levels **A1** to **A6**), and the factor *Visualized room* by the respective visual stimulus components (levels **V1** to **V6**). An integrative survey design covered both the co-presence and the conflicting stimulus paradigms while avoiding a redundant presentation of **AV** congruent stimuli across the paradigms. To limit the total sample to a practicable size, these four factors had to be realized as within-subjects factors. **A** and **V** stimuli were presented first, followed by **A*i*-V*j*** (including **AV**, i.e., $i = j$) stimuli. Within these two test partitions, the stimuli were presented in individually randomized order. By introducing the between-subjects factor *Content*, the total sample was divided into two groups assigned to the music and speech renditions, respectively.

The number of trials within a test sequence corresponds to the number of experimental conditions (factor level combinations). There are two options for

Label	**KH**	**RT**	**KO**	**JC**	**KE**	**GH**
Name	**Konzert-haus**	**Renais-sance Theater**	**Komi-sche Oper**	**Jesus-Christus-Kirche**	**Kloster Eberbach**	**Gewandhaus**
Function	**Concert hall**	**Theatre**	**Opera**	**Church**	**Church**	**Concert hall**
Volume V [m^3]	1899	1903	7266	8079	19539	22202
Size S [m]	12.383	12.392	19.369	20.066	26.934	28.106
Position of receiver (row no./seat no.)	6/8–9	11/178	9/20	3/-	-/-	6/9
Distance receiver—central source D [m]	9.97	9.90	9.46	7.19	15.84	9.84
Absorption coefficient $\alpha_{mean(Sabine)}$	0.18	0.20	0.30	0.17	0.02	0.28
Reverberation time $RT30_{mid}$ [s]	1.29	0.80	1.31	2.81	7.92	2.29
Early Decay Time EDT_{mid} [s]	1.31	0.72	1.17	2.67	8.20	1.99

Table 2.
Geometric and material properties of the selected rooms (taken from [85]). The index mid *refers to the mean of two-octave bands (500 Hz, 1 kHz).*

allocating the trials to the scale items: (a) A long stimulus (ca. 2:00 min, cf. 2.5) is judged by means of the 21 items (2.2); there is just one test sequence. (b) A short stimulus (ca. 6 sec) is judged by means of one feature; the number of test sequences corresponds to the number of features. Option (a) was chosen for the following reasons: (1) In the case of option (b), the comparison of the features, as required by the research project (2.2), would be confounded with the repetition of a stimulus, including greater time intervals, whereas it is not in case of option (a). (2) Short stimuli would run counter to both the context (1.1) and the methodological aim (2.2, 2.3) of the study: artistic renditions are much longer than a few seconds, and—particularly regarding the aesthetic and presence features—responses to very short extracts could not be generalized for entire renditions. (3) To yield valid responses, stimuli must provide enough time and information for judgment formation. Building up an aesthetic impression about very short extracts of an artistic rendition would be hardly possible due to the lack of information about the course of time. Thus, artistically self-contained sections were to be presented at least. Long stimuli provide a greater number and variety of physical events, so that each participant can rely on the individually most helpful cues. (4) In the case of option (a) the decision times vary and are unknown, i.e., within the samples, decision times, as well as causal events and their cues, are pooled. On the one hand, this increases the external validity. On the other hand, it also decreases the internal validity, though, to an acceptable level, since both physical distance and size are constant within each stimulus, and attribution of the estimates to detailed cues or events is not part of the research questions (cf. 1.3).

2.4 Sample

The required sample size was calculated a priori with the aid of the software package G*POWER 3 [85, 86]. Since the groups of the factor *Content* were analyzed separately, only full-factorial repeated measures designs were considered. The

sample size had to be geared to the small 3 × 6 co-presence design. To statistically reveal a relatively small effect size (f = 0.15) at a type I error level of α = 0.05 and a test power of $1-\beta$ = 0.95 while assuming a correlation amongst the repeated measurements of r = 0.6 and an optional nonsphericity correction of ε = 0.7, the minimum sample size per group accounted for n = 38. A total of 114 subjects being affine to music per self-report were initially recruited for the experiment. Subjects were excluded in the following cases (multiple incidences possible):

- Hypoacusis; criterion: audiogram, hearing threshold >20 dB HL at either ear at any of seven tested frequency bands (125 to 8000 Hz), uncompensated by hearing aid (0 subjects).
- Vision deficits; criterion: self-reported deficits, uncompensated by visual aid (0 subjects).
- Red and/or green color blindness; criterion: unpassed Ishihara tests for protanomaly and deuteranomaly (3 subjects).
- Loss of stereopsis; criterion: unpassed contour stereopsis test using the shutter glasses of the projection system (4 subjects).
- Technical incident; failure of saving response data (6 subjects).
- Subjectively untrue responses; criterion: implausible perceptual bias (factor ≥ 5) with reference to visual geometric dimensions (14 subjects, most frequent response: "0 m").

The resultant valid net sample sizes accounted for n = 50 for the music group and for n = 38 for the speech group, comprising 32 female and 56 male voluntary non-experts aged from 21 to 65 years. The frequencies of the participants within the age classes (20s, 30s, 40s, 50s, 60s) amount to f_{abs} = {36; 24; 13; 10; 5}. Participants did not receive incentives.

2.5 Stimuli

As far as possible in a virtual environment, a maximum ecological validity of the stimuli was sought by selecting dedicated performance rooms, artistic content and professional music and speech performers.

Six performance rooms differing in volume (low, medium, high) and average acoustic absorption coefficient (low: $\alpha_{mean(Sabine)} < 0.2$; high: $\alpha_{mean(Sabine)} \geq 0.2$) were selected. Taking into account good speech intelligibility and an accurate perceptibility of the physical room properties (e.g., the visibility of the ceiling height), optimum receiver positions were defined. Based on geometric measures acquired in situ, models of the interior spaces, including the source-receiver-arrangements, were built using the software *SketchUp* (by Google/Trimble) and the plugin *Volume Calculator* (by TGI). The volumes and surface areas of the rooms were then calculated. Standard acoustic measures were taken in situ, in dependence on DIN EN ISO 3382-1 [87]. To corroborate the rooms' selection according to the absorption criterion ex post, Sabine absorption coefficients were calculated from the reverberation times and the geometric properties [88]. The air absorption effect was included; attenuation coefficients were taken from [89]. **Table 2** presents geometric and material properties. Distances were measured directly (i.e., not necessarily in the horizontal plane) from the acoustic center of the central sound source to the

interaural center of the head and torso simulator; they all cover the extrapersonal space. Detailed acoustic measurement reports (research data) are available [90].

The artistic content comprised a musical work and a text, which were chosen to support the perceptibility of the specific room properties by featuring, e.g., impulsivity and sufficient pauses. Two-minute excerpts of Claude Debussy's String Quartet in g minor, op. 10, 2nd movement, and of Rainer Maria Rilke's 1st Duino Elegy were selected. The artistic renditions were audio recorded in the anechoic room of the Technische Universität Berlin.

The performances were presented in the Virtual Concert Hall at Technische Universität Berlin, providing virtual acoustic and visual 3D renditions in rooms. It was particularly designed to meet the methodological requirements (2.1, 2.3), and was completely based on directional binaural room impulse responses (BRIRs) and stereoscopic panoramic images acquired in situ by means of the head and torso simulator *FABIAN* [91, 92]. The stimulus reproduction applied dynamic binaural synthesis by means of an extraaural headset and a semi-panoramic active stereoscopic video projection featuring an effective physical resolution of 4812 × 1800 pixels (**Figure 1**).

The used BRIRs contained the fixed HRTFs of *FABIAN*, hence non-individual HRTFs with regard to the listeners. Experimentation showed that head tracking in connection with non-individual HRTFs improves externalization [93], virtually eliminates front/back confusion, and substantially reduces elevation errors [94]. The auralization system used for this study included head tracking with an angular resolution of 1° and an angular range of ±80° which had to be proved sufficient [95, 96]. It also compensated for spectral coloration [97]. Experimentation also showed that non-individual HpTF compensation, as applied for the present study, outperforms individual HpTF compensation in the specific case of non-individual binaural recordings [98]. System latency was minimized to a level below the perceptual threshold [99]. Cross-fade artifacts were reduced by the applied rendering

Figure 1.
Participant in the Virtual Concert Hall (visual condition: KO).

algorithm fwonder. The system also allowed for the adaption to the participants' individual ITDs [100].

The virtual environment did not provide auditory motion parallax cues by supporting lateral motion interactivity and rendering. This was due to limited in-situ acquisition times in the performance rooms. It would have required measurements at several additional positions of the head and torso simulator, depending on the content-specific minimum audible BRIR grid [101, 102], and thus would have multiplied the expenditure of acquisition time beyond the rooms' availability. However, auditory motion parallax, describing the change in the angular direction of a distant sound source due to the movement of the listener, is assumed to be a supporting cue in absolute distance estimation [103] and known to be a cue in relative depth estimation [104]. Regarding a distance range within the personal space, it was demonstrated by means of a depth discrimination task, and under exclusion of all other distance cues, that auditory motion parallax is exploited by listeners allowing for the perception of distance differences of unknown acoustic stimuli [104]. The cue was shown to be effective for distances between 0.3 and 1.0 m and to be exploitable for lateral head movements within a range of 46 cm. The participants' sensitivity was highest during self-induced motion. Even sensitive subjects did not perceive distance differences corresponding to angular displacements below 3.2°. This value is higher than the minimum audible movement angles (MAMAs) found in previous research (see [105] for an overview). Regarding a distance range of 1 to 10 m, Rumukkainen and colleagues determined the self-translation minimum audible angle (ST-MAA) to be 3.3° by means of 2AFC discrimination tasks without an external reference [106]. Taking into account the absence of external references in the present study and applying the ST-MAA to the nearest sound source used (7.19 m), a concertgoer would remain below perceptual threshold within a lateral moving range of ±41.5 cm, which corresponds to 150% of a typical concert seat's width. Respective lateral movements are normally not observed amongst visitors of classical concerts. Since a relative lateral shift of the listener above the perceptual threshold is a precondition for yielding distance information from the auditory motion parallax cue by triangulation, we do expect neither an appreciable bias nor a deterioration of the accuracy of distance perception introduced by the absence of lateral motion interactivity and rendering.

As a result, the Virtual Concert Hall at Technische Universität Berlin provided almost all relevant auditory cues without major biases (rich-cue condition). Exceptions are the missing supports for (rarely performed and normally small) head orientations around the pitch and roll axes.

The sound pressure level of the virtual rendition was adjusted to the sound pressure level of a live rendition of a string quartet in a real room, which was recorded by the calibrated head and torso simulator. Accounting for the gain of the signal chain and the rooms' STI measures, the level of the scenes' average sound pressure level at the blocked ear canal was L_p = 72.5 dB SPL for a selected *mezzoforte* passage. Likewise, the speech's sound pressure level was adapted to a rendition in a real room and averaged out at L_p = 59.5 dB SPL for a moderate declamatory dynamics stage.

The acquisition of the visual rendering data applied a fixed stereo base, which does not necessarily accord with the participants' individual interpupillary distances (IPDs). Respective differences might potentially bias the individual distance and room size perception. To date, experimentation has shown inconsistent effects of the variation of IPD differences on distance perception (see [46] for a review). Most studies cannot be translated into the present study, since they investigated maximum target distances of 1 m and/or used simple numerically modeled objects/environments. Moreover, results differ regarding the significance, the size and/or the direction of the effects. This is apparently due to different rendering

technologies (stereoscopic projection, HMD, CAVE), stages of virtualization (mixed reality, virtual reality), target distances (personal space, action space), simulated objects/environments (simple graphic objects, shapes, persons in hallways), and measurement protocols (triangulated distance estimation, blind walking, visual alignment, verbal estimation) [107–113]. Few experiments investigated distances roughly similar to those used in the present study (about 7 to 16 m). While Willemsen and colleagues did not observe a significant effect of IPD individualization on distance judgments [114], a large variation of the stereo base (0 to 4 times the IPD) showed significant effects on both distance and size judgments: Greater stereo bases resulted in perceptually closer and smaller objects [115]. However, relevance for the descriptive measures, effect sizes and significances of the present study is given rather by the expected value and distribution of the IPD differences than by their individual values. Anthropometric data of the German resident population, from which the sample was drawn, state median IPDs of 61 mm (male persons) and 60 mm (female persons) within the age range of 18 to 65 years [116]. Since the values do nearly exactly meet the stereo base of the target acquisition (60 mm), a substantial collective perceptual bias is unlikely to occur.

Limitations of the visual rendering pertain to the field of view (161° × 56°), which should at least not affect distance perception [59, 117]; the angular resolution (2.1 arcmin), which might affect distance perception [57]; the fixed single focal plane in stereoscopy providing an invariant accommodation cue, so that the connection between convergence and accommodation is suspended [45]; and an undersized luminance of the projection. Data projectors could not provide the luminance and the contrast of the real scenes, especially in connection with shutter glasses. Thus, the luminances of the scenes were fitted into the projectors' dynamic range while maintaining compressed relations of the luminances. Scene luminances were calculated from exposure time, aperture, and ISO arithmetic film speed of correctly exposed photographs of a centrally placed and vertically oriented 18% gray card according to the additive system of photographic exposure (APEX). The average loss of the luminance value *Bv* introduced by the projection and shutter glasses was 2.88. The average scene luminance L_v of the gray cards amounted to 0.82 cd/m^2. Detailed information regarding room acquisition, content production, and stimulus reproduction for the Virtual Concert Hall was published separately [118].

Since electronic media transform both the physical stimuli and their perception, the replacement of natural by mediatized stimuli for serious experimental purposes demands the knowledge of the perceptual influences of the applied mediatizing system, as also pointed out by [16, 21]. The rendering technique of the Virtual Concert Hall was shown to provide perceptually plausible auralizations [119]. Specifically, the Virtual Concert Hall at Technische Universität Berlin was subjected to a test of auditory-visual validation by comparing a real scene and the correspondent virtual scene [38]. Amongst others, it yielded nearly equal loudness judgments of the real and the virtual environment, whereas the virtual environment—apparently due to the dark surrounding—was perceived slightly brighter than the respective real environment. The virtualization also generally lowered the perceived source distance and the perceived size of a real room—mainly due to the visual rendering. The mere auditory underestimation of source distance and room size introduced by the virtualization amounted only to 6.6 and 1.9%, respectively. The biases are considered in the discussion section.

2.6 Procedure

Each participant ran through the test procedure individually. The procedure lasted about 3 hours and 10 minutes, and comprised color vision and stereopsis

tests, audiometry, a socio-demographic questionnaire, a privacy agreement, the clarification of the questionnaire, the measurement of the individual inter-tragus distance (necessary for the technical adaption to the individuals' ITDs), cabling, a familiarization sequence, and the actual test runs, inclusive of self-imposed breaks.

2.7 Data analysis

Arithmetic means standard deviations (**Tables 11** and **12**) and standard errors were calculated for all combinations of factor levels. The means were plotted against the combinations. According to the test design (2.3), the co-presence paradigm required 3 × 6 repeated measures analyses of variance (rmANOVA), the conflicting stimulus paradigm 6 × 6 rmANOVA for either level of *Content*. *Content* was not regarded as a factor for analysis because it was not covered by the RQs, and the quantification of the proportions according to RQs 3–5 were to be made possible separately for both music and speech. Kolmogorov-Smirnov tests indicated that the assumption of normally distributed error components was met with the exceptions of source distance under the conditions speech **A0**-**V5** (KS-Z = 1.390, p = 0.042) and speech **A6**-**V3** (KS-Z = 1.442, p = 0.031), and of room size under the conditions speech **A0**-**V5** (KS-Z = 1.500, p = 0.022), speech **A5**-**V0** (KS-Z = 1.759, p = 0.004), and music **A4**-**V5** (KS-Z = 1.428, p = 0.034). The minor violations concerning 4.8% of the conditions were deemed tolerable because of the robustness of the rmANOVA. Mauchly's sphericity tests indicated a significant violation of the sphericity assumption in both the 3 × 6 and the 6 × 6 analyses, which was compensated for by correcting the degrees of freedom using Greenhouse-Geisser estimates. To answer RQs 1 and 2, an orthogonal set of planned main contrasts (reverse Helmert) was calculated: Simple contrast **V** vs. **A**; combined contrast **VA** vs. {**V**, **A**}. To allow different approaches to effect size comparison, partial eta squared η_P^2, classical eta squared η^2, and generalized eta squared η_G^2 [120, 121] were reported for the omnibus tests. Because of RQs 3–5, and taking advantage of the commensurability of the factors *Auralized room* and *Visualized room* of the conflicting stimulus design, the η^2 effect sizes were particularly reported as indicators for the proportional influence of the acoustic room properties, the visual room properties, and their interaction on the geometric features. To allow their direct comparison in a simplified manner, the net effect sizes (the proportions of the explained variance) given by $\eta_{X(net)}^2 = \eta_X^2/\left(\eta_A^2 + \eta_V^2 + \eta_{A\times V}^2\right)$ were also reported. Based on Cohen's f ([122], p. 281), which was calculated from η^2 ([123], p. 7), the effect sizes were classified as small, medium or large.

3. Results

3.1 Perceived source distance

3.1.1 Co-presence paradigm

Source distance showed significant main and interaction effects of *Domain* and *Room* for both music (**Table 3**) and speech (**Table 4**). Effects were large for *Room* and medium size for *Domain* and *Domain* × *Room*. The mean distance estimates were generally lower for speech than for music, and the range of the mean estimates introduced by the factor *Domain* was lower for the low-absorbent (wet) and higher for the high-absorbent (dry) rooms, even though it was not hypothesized or tested (**Figures 2** and **3**; **Tables 11** and **12**).

S. o. V.	*SS*	df_{adj}	*MS*	*F*	*p*	η^2	*f*	η^2_G	η^2_P	1-β
Domain	1521.061	1.782	853.503	36.965	<0.001	0.086	0.306	0.122	0.430	>0.999
Error (*Domain*)	2016.285	87.325	23.090							
Room	4610.610	3.845	1199.166	137.464	<0.001	0.260	0.593	0.296	0.737	>0.999
Error (*Room*)	1643.487	188.397	8.724							
Domain × *Room*	597.939	7.113	84.059	9.593	<0.001	0.034	0.187	0.052	0.164	>0.999
Error (*D.* × *R.*)	3054.229	348.552	8.763							

Table 3.
Results of the rmANOVA for perceived source distance $\hat{D}$ *(music, co-presence paradigm).*

S. o. V.	*SS*	df_{adj}	*MS*	*F*	*p*	η^2	*f*	η^2_G	η^2_P	1-β
Domain	655.466	1.683	389.381	23.350	<0.001	0.058	0.248	0.073	0.387	>0.999
Error (*Domain*)	1038.621	62.284	16.676							
Room	1712.901	3.387	505.745	41.676	<0.001	0.152	0.423	0.171	0.530	>0.999
Error (*Room*)	1520.729	125.315	12.135							
Domain × *Room*	639.600	5.709	112.027	10.836	<0.001	0.057	0.245	0.072	0.227	>0.999
Error (*D.* × *R.*)	2183.952	211.245	10.338							

Table 4.
Results of the rmANOVA for perceived source distance $\hat{D}$ *(speech, co-presence paradigm).*

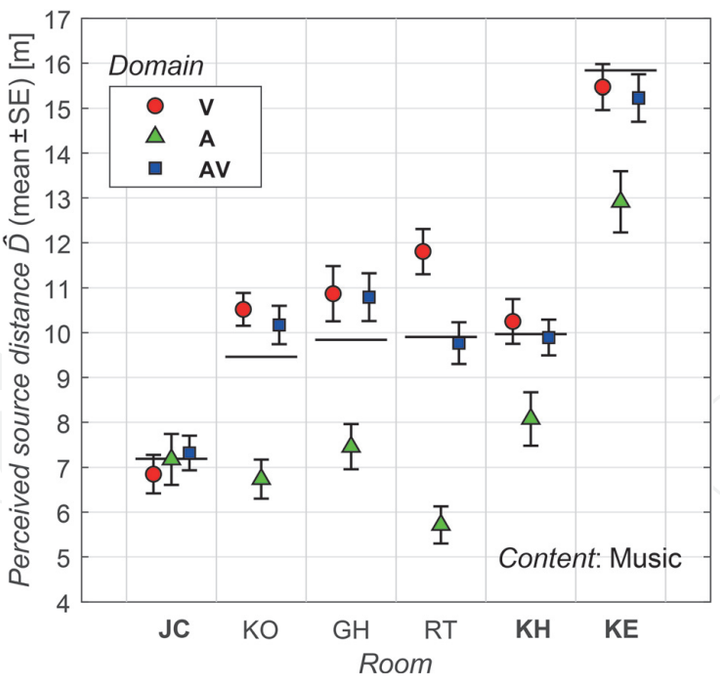

Figure 2.
Means (markers) and standard errors (bars) of perceived source distance $\hat{D}$ *against factor levels of* Room *and* Domain *for music. Horizontal lines indicate the particular physical source distance D within each room. Bold labels indicate low-absorbent rooms.*

Regarding RQ 1, a priori main contrasts indicate that the mean estimates at level **V** were considerably higher than those at level **A**. The mean differences account for 2.95 m (music), $F(1,49) = 52.910$, $p < 0.001$, $\eta^2_P = 0.519$, and for 2.38 m (speech), $F(1,49) = 32.712$, $p < 0.001$, $\eta^2_P = 0.469$. This is also consistent on a descriptive basis

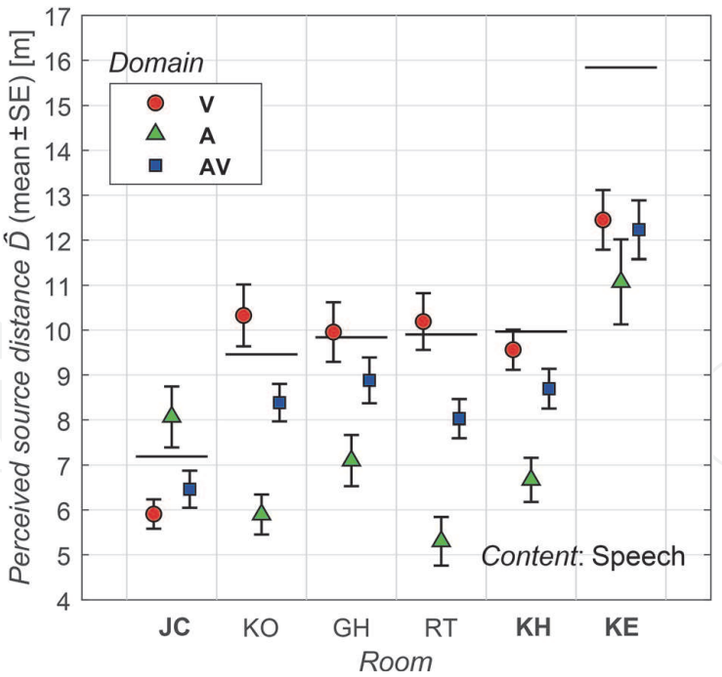

Figure 3.
Means (markers) and standard errors (bars) of perceived source distance $\hat{D}$ *against factor levels of* Room *and* Domain *for speech. Horizontal lines indicate the particular physical source distance D within each room. Bold labels indicate low-absorbent rooms.*

across all rooms except JC, which involves the smallest physical distance (s = 7.19 m) and shows a lower mean estimate under the **V** than under the **A** condition for both music and speech. Looking at RQ 2, the mean estimates at level **AV** were higher than the average of the mean estimates at levels **A** and **V**. The mean differences accounted for 1.04 m in the music group (a priori main contrast), $F(1,49)$ = 13.141, p = 0.001, η_P^2 = 0.211, and 0.24 m in the speech group (contrast not significant). The **AV** mean estimates were located at 85% of the range between the mean estimates at levels **V** and **A** in the music group and at 60% in the speech group.

Looking at the accuracy of the estimates, the mean estimates differed from the mean physical source distance by −2.36 m (−22.7%) at level **A**, +0.59 m (+5.7%) at level **V**, and +0.16 m (+1.5%) at level **AV** in the music group, and by −3.02 m (−29.1%) at level **A**, −0.63 m (−6.1%) at level **V**, and −1.59 m (−15.3%) at level **AV** in the speech group. Overall, the physical distances were met best by the estimates at level **AV** in the music group, and by the estimates at level **V** in the speech group.

3.1.2 Conflicting stimulus paradigm

Auralized room and *Visualized room* showed significant main effects on source distance for both music (**Table 5**) and speech (**Table 6**), however, no significant interaction effect. Effects of *Auralized room* were of small size, whereas effects of *Visualized room* were classified as large. Regarding music, $\eta^2_{A(net)}$ = 7% of the proportion of the explained variance (see 2.7) arose from *Auralized room*, $\eta^2_{V(net)}$ = 91% from *Visualized room*. Under the speech condition, the proportions accounted for 11% (*Auralized room*) and 88% (*Visualized room*).

Figures 4 and **5** show the generally lower mean distance estimates for the speech by trend. The figures also illustrate the ranges of the mean estimates. The average range of mean estimates caused by *Auralized room* was 1.69 m, while the range caused by *Visualized room* accounted for 5.74 m. The range of the physical source

S. o. V.	*SS*	df_{adj}	*MS*	*F*	*p*	η^2	*f*	η^2_G	η^2_P	1-β
Auralized room	469.724	5.000	93.945	13.143	<0.001	0.017	0.133	0.023	0.211	>0.999
Error (*A. room*)	1751.252	131.608	13.307							
Visualized room	6256.608	2.602	2404.324	105.444	<0.001	0.233	0.551	0.238	0.683	>0.999
Error (*V. room*)	2907.446	127.509	22.802							
A. room × *V. room*	134.833	13.677	9.858	1.566	0.086	0.005	0.071	0.007	0.031	0.868
Error (*A. r.* × *V. r.*)	4219.961	670.192	6.297							

Table 5.
Results of the rmANOVA for perceived source distance $\hat{D}$ *(music, conflicting stimulus paradigm).*

S. o. V.	*SS*	df_{adj}	*MS*	*F*	*p*	η^2	*f*	η^2_G	η^2_P	1-β
Auralized room	375.912	1.667	225.526	9.314	0.001	0.023	0.153	0.028	0.201	0.951
Error (*A. room*)	1493.259	61.672	24.213							
Visualized room	2936.460	2.724	1077.931	48.375	<0.001	0.178	0.465	0.183	0.567	>0.999
Error (*V. room*)	2245.993	100.794	22.283							
A. room × *V. room*	31.620	12.609	2.508	0.531	0.902	0.002	0.044	0.002	0.014	0.317
Error (*A. r.* × *V. r.*)	2203.942	466.540	4.724							

Table 6.
Results of the rmANOVA for perceived source distance $\hat{D}$ *(speech, conflicting stimulus paradigm).*

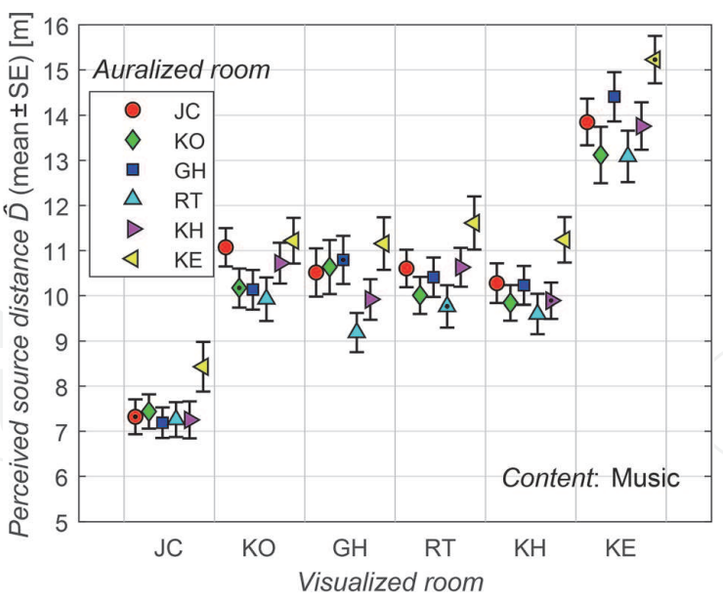

Figure 4.
Means (markers) and standard errors (bars) of perceived source distance $\hat{D}$ *against factor levels of* Auralized room *and* Visualized room *for music. Dots within markers indicate acoustic-visual congruency.*

distance was 8.65 m. As a rule, the auralized room KE led to a maximal mean estimate and the auralized room RT to a minimal mean estimate within each visualized room. In turn, the visualized room KE led to a maximal mean estimate and the visualized room JC to a minimal mean estimate within each auralized room. The mean estimates do not indicate that acoustic-visual congruency as such yielded maximal, minimal or especially accurate mean distance estimates.

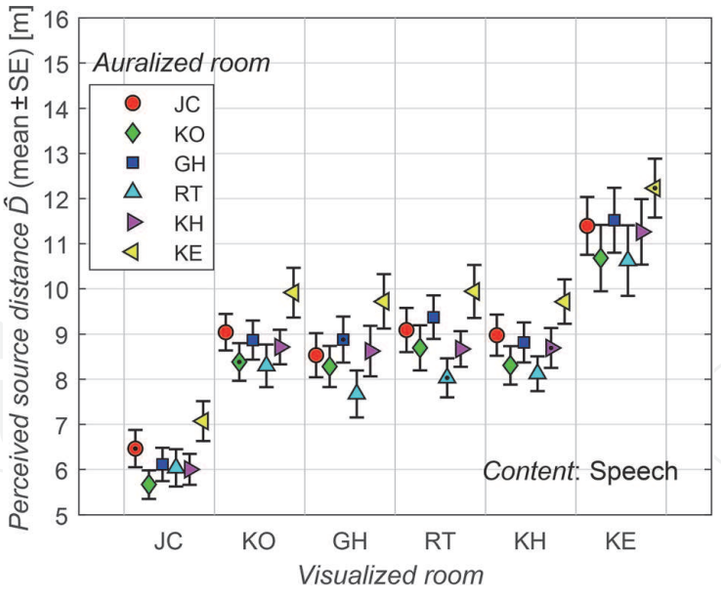

Figure 5.
Means (markers) and standard errors (bars) of perceived source distance $\hat{D}$ *against factor levels of* Auralized room *and* Visualized room *for speech. Dots within markers indicate acoustic-visual congruency.*

3.2 Perceived room size

3.2.1 Co-presence paradigm

Room size showed significant main and interaction effects of *Domain* and *Room* for both music (**Table 7**) and speech (**Table 8**). Effects were of large size for *Domain* (music) and *Room* and of medium size for *Domain* (speech) and *Domain* ×

S. o. V.	*SS*	df_{adj}	*MS*	*F*	*p*	η^2	*f*	η^2_G	η^2_P	1-β
Domain	10148.965	1.651	6145.611	70.421	<0.001	0.115	0.361	0.180	0.590	>0.999
Error (*Domain*)	7061.808	80.919	87.270							
Room	28109.442	3.650	7701.183	226.890	<0.001	0.319	0.685	0.379	0.822	>0.999
Error (*Room*)	6070.632	178.851	33.942							
Domain × *Room*	3733.981	7.522	496.424	21.814	<0.001	0.042	0.210	0.075	0.308	>0.999
Error (*D.* × *R.*)	8387.358	315.667	26.570							

Table 7.
Results of the rmANOVA for perceived room size $\hat{S}$ *(music, co-presence paradigm).*

S. o. V.	*SS*	df_{adj}	*MS*	*F*	*p*	η^2	*f*	η^2_G	η^2_P	1-β
Domain	6484.837	1.513	4285.200	42.093	<0.001	0.082	0.299	0.131	0.532	>0.999
Error (*Domain*)	5700.224	55.992	101.803							
Room	26573.103	2.994	8875.557	165.259	<0.001	0.337	0.713	0.383	0.817	>0.999
Error (*Room*)	5949.496	101.789	58.449							
Domain × *Room*	3000.770	6.785	442.292	19.791	<0.001	0.038	0.199	0.065	0.348	>0.999
Error (*D.* × *R.*)	5610.150	251.030	22.349							

Table 8.
Results of the rmANOVA for perceived room size $\hat{S}$ *(speech, co-presence paradigm).*

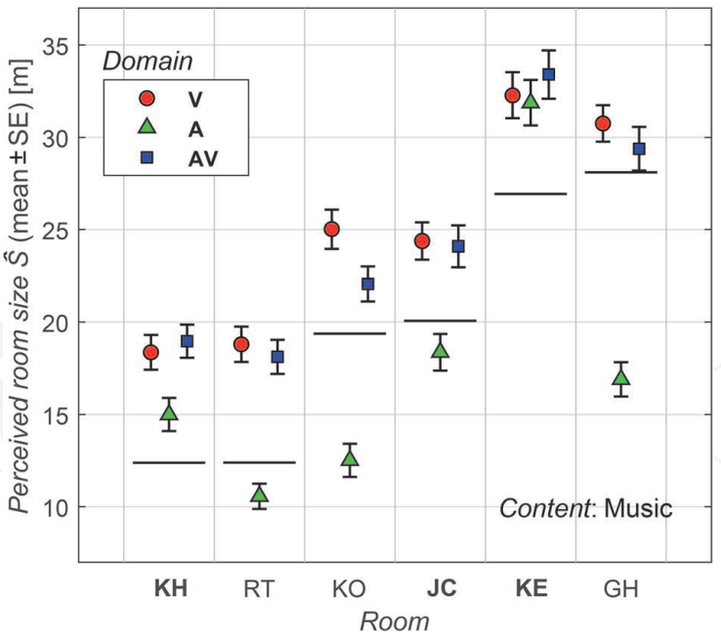

Figure 6.
Means (markers) and standard errors (bars) of perceived room size $\hat{S}$ *against factor levels of* Room *and* Domain *for music. Horizontal lines indicate the particular physical room size S of each room. Bold labels indicate low-absorbent rooms.*

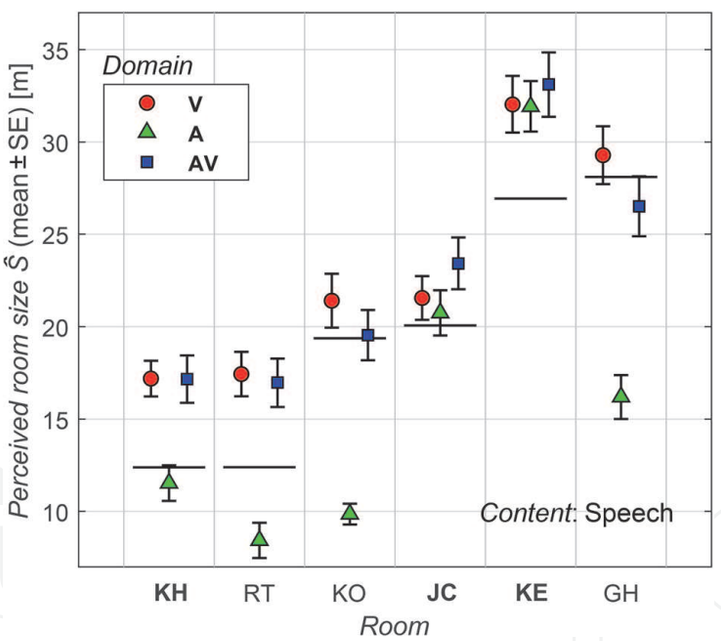

Figure 7.
Means (markers) and standard errors (bars) of perceived room size $\hat{S}$ *against factor levels of* Room *and* Domain *for music. Horizontal lines indicate the particular physical room size S of each room. Bold labels indicate low-absorbent rooms.*

Room. The mean size estimates were slightly lower for speech than for music by trend (**Figures 6** and 7).

Regarding RQ 1, a priori contrasts indicated that the mean estimates at level **V** were considerably higher than those at level **A**. The mean differences account for 7.40 m (music), $F(1,49) = 97.748$, $p < 0.001$, $\eta_P^2 = 0.666$, and for 6.71 m (speech), $F(1,49) = 51.457$, $p < 0.001$, $\eta_P^2 = 0.582$. Looking at RQ 2, a priori contrasts showed that the mean estimates at level **AV** were higher than the average of the mean estimates at levels **A** and **V**. The mean differences accounted for 3.11 m (music),

$F(1,49) = 32.124$, $p < 0.001$, $\eta_P^2 = 0.396$, and 2.99 m (speech), $F(1,49) = 24.933$, $p < 0.001$, $\eta_P^2 = 0.403$. The **AV** estimates were located at 92% of the range between the mean estimates at levels **V** and **A** in the music group and at 94% in the speech group.

As with source distance, the range of the mean room size estimates introduced by the factor *Domain* was lower for the low-absorbent (wet) and higher for the high-absorbent (dry) rooms, even though this was not hypothesized or tested.

Accuracies were generally low regardless of the level of *Domain*. The mean room size estimates differed from the mean physical room size by −2.00 m (−10.0%) at level **A**, +5.47 m (+27.5%) at level **V**, and +4.86 m (+24.5%) at level **AV** in the music group, and by −3.08 m (−15.5%) at level **A**, +3.72 m (+18.7%) at level **V**, and +3.34 m (+16.8%) at level **AV** in the speech group. Overall, the physical sizes were generally best approximated by the estimates at level **A**. Specifically, in low-absorbent rooms (KH, JC, KE) and the small dry room (RT), physical room sizes were best approximated by the estimates at level **A**, whereas in medium- and large-sized dry rooms (KO, GH) they were best approximated by the estimates at levels **AV** and **V**.

3.2.2 Conflicting stimulus paradigm

Auralized room and *Visualized room* showed significant main effects on room size for both music (**Table 9**) and speech (**Table 10**), however, no significant interaction effect. Effects of *Auralized room* were of small size, whereas effects of *Visualized room* were classified as large. Regarding music, $\eta^2_{A(net)} = 9\%$ of the proportion of the explained variance (see 2.7) arose from *Auralized room*, $\eta^2_{V(net)} = 90\%$

S. o. V.	*SS*	df_{adj}	*MS*	*F*	*p*	η^2	*f*	η^2_G	η^2_P	1-β
Auralized room	3275.179	2.048	1599.570	25.911	<0.001	0.024	0.156	0.031	0.346	>0.999
Error (*A. room*)	6193.742	100.329	61.734							
Visualized room	32107.238	3.203	10025.415	110.275	<0.001	0.233	0.551	0.239	0.692	>0.999
Error (*V. room*)	14266.617	156.927	90.913							
A. room × *V. room*	375.257	12.004	31.262	1.344	0.189	0.003	0.052	0.004	0.027	0.754
Error (*A. r.* × *V. r.*)	13678.450	588.172	23.256							

Table 9.
Results of the rmANOVA for perceived room size $\hat{S}$ *(music, conflicting stimulus paradigm).*

S. o. V.	*SS*	df_{adj}	*MS*	*F*	*p*	η^2	*f*	η^2_G	η^2_P	1-β
Auralized room	3799.130	1.446	2626.800	11.517	<0.001	0.026	0.162	0.030	0.237	0.968
Error (*A. room*)	12205.465	53.513	228.084							
Visualized room	23087.978	2.228	10363.307	54.821	<0.001	0.155	0.429	0.160	0.597	>0.999
Error (*V. room*)	15582.628	82.431	189.039							
A. room × *V. room*	185.804	7.540	24.642	0.662	0.716	0.001	0.035	0.002	0.018	0.296
Error (*A. r.* × *V. r.*)	10382.097	278.982	37.214							

Table 10.
Results of the rmANOVA for perceived room size $\hat{S}$ *(speech, conflicting stimulus paradigm).*

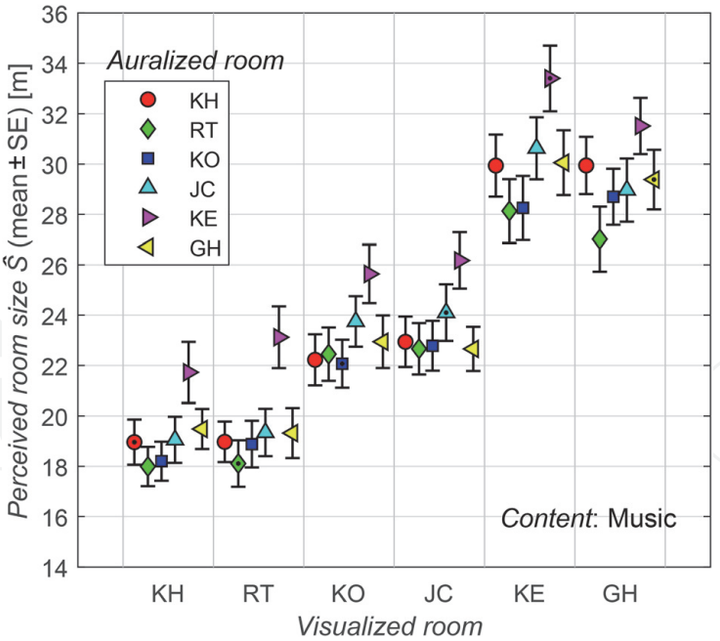

Figure 8.
Means (markers) and standard errors (bars) of perceived room size $\hat{S}$ *against factor levels of* Auralized room *and* Visualized room *for music. Dots within markers indicate acoustic-visual congruency.*

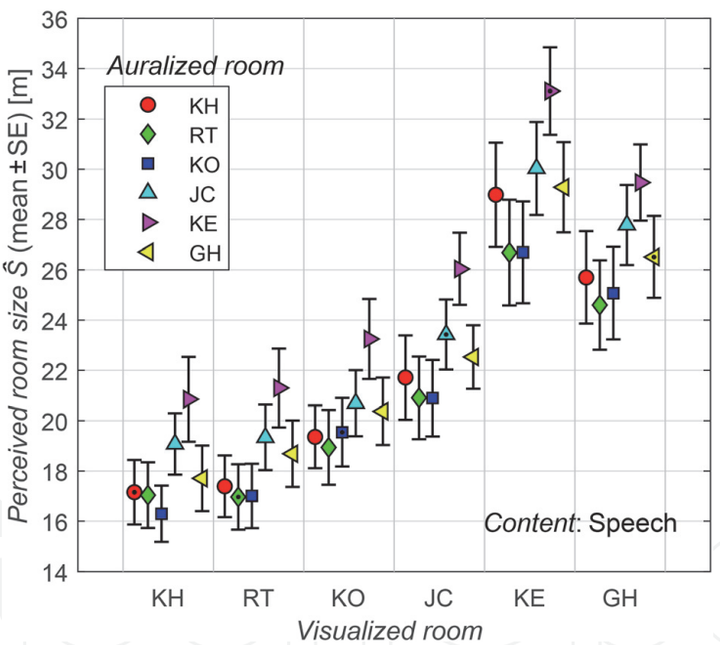

Figure 9.
Means (markers) and standard errors (bars) of perceived room size $\hat{S}$ *against factor levels of* Auralized room *and* Visualized room *for speech. Dots within markers indicate acoustic-visual congruency.*

from *Visualized room*. Under the speech condition, the proportions accounted for 14% (*Auralized room*) and 85% (*Visualized room*).

Figures 8 and **9** show the generally lower mean room size estimates for the speech by trend. The figures also illustrate the ranges of the mean estimates. The average range of mean estimates caused by *Auralized room* was 4.54 m, the range caused by *Visualized room* accounted for 10.99 m. The range of the physical room size was 15.72 m. As a rule, the auralized room KE led to a maximal mean estimate and the auralized room RT to a minimal mean estimate within each visualized room. In turn, the visualized room KE led to a maximal mean estimate and the visualized room KH mostly to a minimal mean estimate within each auralized room. The mean

estimates do not indicate that acoustic-visual congruency as such yielded maximal, minimal or especially accurate mean size estimates.

4. Discussion

4.1 Presence of auralized and visualized rooms

Most of the results apply likewise to both egocentric distance and room size estimation. RQ 1 asked for the difference between the modalities as such. Mean estimates across the rooms based only on visual information significantly and considerably exceeded those based only on acoustic information, specifically by about a fourth of the mean physical property in the case of distance and by about a third in the case of size. Hence, $H1_1$ can be accepted and might be reformulated directionally ($H1_1$: $\mu_A < \mu_V$) for future experimentation. Regarding egocentric distance estimation, the finding is plausible in principle given the reported compression of distance perception in real acoustic environments [27, 28, 31–33] and virtual acoustic environments [32, 34–36]. However, it does not agree with [36], who observed a compressed perception of visual distances between 1.5 and 5.0 m, or with [18], who used nearly the same auralization system in connection with smaller distances (1.93–5.88 m) and a restricted visualization. Though the general finding $\hat{D}_V > \hat{D}_A$ does also not accord with the finding of [38] under the virtual environment condition ($\hat{D}_V < \hat{D}_A$), the exceptional observation at the smallest physical distance (D = 7.19 m, room JC) does. This is likely to be due to the same physical distance used in [38], indicating that the general finding might be confined to physical distances greater than about 8 m. However, checking the acceptance of inference from virtual rooms to real rooms for the music content by multiplying the mean estimates of the present study by the reality-to-virtuality ratios (*RVR*s) of the mean estimates of [38] ($RVR_{\text{distance,A}}$ = 1.071, $RVR_{\text{distance,V}}$ = 1.318; $RVR_{\text{size,A}}$ = 1.019, $RVR_{\text{size,V}}$ = 1.236) allows the findings $\hat{D}_V > \hat{D}_A$ and $\hat{S}_V > \hat{S}_A$ to be transferred from virtual scenes to corresponding real scenes without the persistence of the aforesaid scene-specific exception.

4.2 Basic mode of perception

Regarding RQ 2, there is evidence that the basic mode of perception (processing of single- vs. multi-domain stimuli) as such alters perceptual estimates of geometric dimensions in virtual rooms. Mean estimates based on acoustic-visual stimuli did not equal the average of the mean estimates based on either only acoustic or only visual stimuli. Rather, mean estimates of source distance under the acoustic-visual condition (with acoustic-visually congruent stimuli) were located at 85% (music) of the range between the mean estimates of the levels **A** and **V**, mean estimates of room size at 92% (music) and 94% (speech), indicating that under the multi-domain condition visual information was weighted significantly higher than acoustic information. Though the distance estimation of the speech performance did not show a significant effect of perceptual mode, the mean estimates still accounted for 60% of the range between the mean estimates at levels **A** and **V**. When loading the mean estimates with the above-mentioned compensation factors, the percentages concerning music changed from 85% to 84% for source distance and from 92–86% for room size. Hence, the finding on RQ 2 may be transferred to reality in principle.

4.3 Properties of auralized and visualized rooms

Considering the multi-domain mode of perception and applying the conflicting stimulus paradigm, the distance and size estimates depended significantly on both the acoustic and the visual properties of the stimuli (RQs 3 and 4). Generally, about 89% of the explained variance arose from the entire visual and 10% from the entire acoustic information provided by the virtual environment. For both egocentric distance and room size perception, acoustic information showed a slightly greater proportion of explained variance under the speech than under the music condition.

In accordance with the MLE modeling of auditory-visual integration in principle, the acoustic and visual proportions of the explained variance appear to vary strongly according to the availability and, respectively, the richness of the cues in the particular domains: A preliminary experiment under substantially restricted visualization conditions (reduced field of view, reduced spatial resolution, still photographs instead of moving pictures, no maximal acoustic-visual congruency due to visible loudspeakers as sound sources) and non-restricted auralization conditions (identical auralization system) yielded a reversed order of proportions of the explained variance (cf. 2.7), which amounted to 33% for factor *Visualized room*, and 66% for factor *Auralized room* ([18], p. 392).

Against the background of the prevalent term *auditory-visual interaction* (or similar) it is remarkable, that at least no *statistical* interaction effect of the acoustic and the visual stimulus properties on egocentric source distance and room size perception was found that was significant (1.3, RQ5). Looking at perceived geometric dimensions as supramodal unified features specifying spatial notions, both acoustic and visual properties, and therefore both the auditory and the visual modalities, appear to contribute (regardless of variable weights) directly to the values of these features, and no interaction (non-additive) effects appear to complicate this straightforward principle. Hence, the modeling of auditory-visual integration of distance and room size perception will not have to include non-additive effects for the time being.

Since the involved modalities and the mode of perception were constant across all factor levels, it may be assumed that VR-induced biases apply likewise to all factor levels of the conflicting stimulus paradigm and their combinations. Hence, the findings on RQs 3 to 5, i.e., the inferential statistics and the η^2-based proportional accounts for the estimates, may be transferred from virtuality to reality in principle. At the descriptive level, the estimates might again be compensated for virtualization by loading them with $RVR_{\text{distance,AV}} = 1.284$ and $RVR_{\text{size,AV}} = 1.191$, respectively [38].

4.4 Complex independent variables and interfering factors

Within the test design, the presence and properties of the acoustic and visual domains were varied to experimentally dissociate the auditory and the visual modalities. Because this variation was categorical, i.e., comprising the entire *environmental* conditions of the scenes instead of either mere *distance* or mere *room size* cues, the results may be transferred to the perceptual modalities hearing and vision as such—at least for closed spaces, and within the boundaries of generalization given by the content types, rooms, and samples. Auditory-visual distance perception may in principle be influenced not only by physical distance, but by any structural (room size, room shape) and material properties that affect those acoustic cues (1.2) that are also affected by physical distance (cf. [124]). Since the domain

proportions found in the present study cannot directly be compared to the weights determined in [79], which are based on mere distance-related cues, those interfering factors had to be experimentally dissociated and, where applicable, included in physical-perceptual models of auditory-visual distance perception.

4.5 Additional observations

There were some additional results on factors and measures which were not explicitly asked for by the RQs:

a. Both egocentric distance and room size mean estimates, regardless of whether based on acoustic, visual or acoustic-visual stimuli, were obviously lower for speech than for music (though this was not hypothesized or tested, see 2.7). Hence, there is a reason for hypothesizing an influence of content type. This might be due to differences between music and speech regarding, e.g., the bandwidth and energy distribution of the frequency spectra carrying spatial information, perceptual filtering and processing, receptiveness, and/or experiential geometric situations (non-mediatized speech is normally received from lower distances and within smaller rooms than non-mediatized music).

b. Both the non-significant interaction effect and the particular mean estimates in the experiment according to the conflicting stimulus paradigm indicated that acoustic-visual (mainly spatial) congruency of the stimulus properties did not lead to minimum, maximum or especially accurate mean estimates. This observation is not apt to constitute a general hypothesis, since congruency might play a greater role by contrast with a greater range of the incongruencies (e.g., further-away sound sources) or a greater number of incongruent properties (e.g., including incongruent content).

c. Egocentric distance mean estimates were most accurate under the acoustic-visual (music) and visual (speech) condition; the room size mean estimates, which were generally inaccurate, likely due to the lack of the visual rendering of the rooms' rear part, were most accurate under the acoustic condition. In contrast to previous studies [32, 36], regardless of general under- or overestimations of the geometric properties ($\hat{D}/D \neq 1$) under the acoustic-visual condition, neither an *increasing* underestimation nor an *increasing* overestimation was conspicuous, rather $\hat{D}/D \approx$ const.

d. Looking at the conflicting stimulus paradigm, the minimum and maximum mean estimates of both source distance and room size did not consistently correspond to the minimum and maximum physical distances and sizes. *Perceived source distance* and *perceived room size* were each influenced by the physical source distance, the physical room size and potentially other properties of the virtual scenes.

e. Because mean estimates based on purely acoustic stimuli were generally higher in low-absorbent than in high-absorbent rooms (cf. [18]), the range of mean estimates introduced by the factor *Domain* was also generally smaller in low-absorbent rooms. This caused the respective mean estimates under the

acoustic condition to be more consistent with—and in the case of room size, even more accurate than—those under the visual and acoustic-visual conditions. Therefore, when visual information is unavailable, perception may exploit the greater amount of acoustic information provided by low-absorbent rooms to improve the accuracy of room size perception. Acoustic absorption may influence not only the values but also the availability and/or acuity of auditory cues (cf. 1.2).

Observations (d) and (e) and differences between the studies regarding domain proportions (4.3) give reason to hypothesize that structural and material properties of rooms influence distance perception. Thus, an additional experimental dissociation of the factors physical source distance, physical room size, and acoustic absorption (all else being equal) might be instructive. Furthermore, more detailed physical factors affecting both the acoustic and the visual domain might be disentangled (primary structures, secondary structures, materials). Because of the trade-off between the requirement of ecological stimulus validity and the costs of stimulus production, it might be worth investigating the moderating effects of certain aspects of virtualization (direct rendering, stereoscopy, visually moving persons). In the future, one major aim of research into the perception of geometric properties might be the connection of the modeling of internal mechanisms and the physical-perceptual modeling.

5. Conclusion

The influence of the presence as well as of the properties of acoustic and visual information on the perceived egocentric distance and room size was investigated applying both a co-presence and a conflicting stimulus paradigm. Constant music and speech renditions in six different rooms were presented using dynamic binaural synthesis and stereoscopic semi-panoramic video projection. Experimentation corroborated that perceptual mean estimates of geometric dimensions based on only visual information considerably exceeded those based on only acoustic information in general. However, the perceptual mode as such (single- vs. multi-domain stimuli) altered the perceptual estimates of geometric dimensions: Under the acoustic-visual condition with acoustic-visually congruent stimuli, the presence of visual geometric information was generally given more weight than the presence of acoustic information. While the egocentric distance estimation under the acoustic-visual condition did not tend to be compressed for music, it did for speech. When only acoustic stimuli were available, the greater amount of acoustic information provided by low-absorbent rooms appeared to be perceptually exploited to improve the accuracy of room size perception. Within the multi-domain mode of perception involving 30 acoustic-visually incongruent and 6 congruent stimuli, auditory-visual estimation of geometric dimensions in rooms relied about nine-tenths on the variation of visual, about one-tenth on the variation of acoustic properties, and negligibly on the interaction of the variation of the particular properties. Both the auditory and the visual sensory systems contribute to the perception of geometric dimensions in a straightforward manner. The observation of generally lower estimates for speech than for music needs to be corroborated and clarified. Further experimentation dissociating the factors source distance, room size, and acoustic absorption (all else being equal) is needed to clarify their particular influence on auditory-visual distance and room size perception.

Ethics statement

According to the funding institution (Deutsche Forschungsgemeinschaft) an ethical approval is not required, since the respective indications do not apply [125]. The study was conducted under the ethical principles of the appropriate national professional society (Deutsche Gesellschaft für Psychologie) [126].

Acknowledgements

This work was carried out as a part of the project "Audio-visual perception of acoustical environments", funded by the Deutsche Forschungsgemeinschaft (DFG MA 4343/1-1) within the framework of the research unit SEACEN, coordinated by Technische Universität Berlin and Rheinisch-Westfälische Technische Hochschule Aachen, Germany. We thank the staff of the performance rooms for their friendly cooperation, the Berlin Budapest Quartet (Dea Szücs, Éva Csermák, Itamar Ringel, Ditta Rohmann) and actress Ilka Teichmüller for their performances, Alexander Lindau, Fabian Brinkmann, and Vera Erbes for the in-situ acquisition of the rooms' acoustic and visual properties, Mina Fallahi for the geometric picture editing, Annika Natus, Alexander Haßkerl, and Shamir Ali-Khan for the 3D video shooting and post-production, and all test participants. Finally, we thank the two anonymous reviewers for critically reading the manuscript and suggesting substantial improvements.

Conflict of interest

The authors have no conflict of interest to declare.

A. Appendix

			V room						
Measure	***Content***	***A room***	**off**	**GH**	**JC**	**KH**	**KO**	**RT**	**KE**
Mean	Music (n = 50)	off	—	10.87	6.85	10.25	10.52	11.80	15.47
		GH	7.46	10.79	7.19	10.23	10.13	10.41	14.41
		JC	7.18	10.51	7.32	10.28	11.07	10.60	13.85
		KH	8.07	9.92	7.25	9.89	10.72	10.63	13.76
		KO	6.74	10.63	7.44	9.84	10.17	10.01	13.12
		RT	5.71	9.19	7.26	9.60	9.92	9.77	13.09
		KE	12.91	11.16	8.43	11.24	11.22	11.61	15.23
	Speech (n = 38)	off	—	9.96	5.91	9.56	10.33	10.19	12.45
		GH	7.10	8.88	6.11	8.82	8.87	9.37	11.52
		JC	8.07	8.53	6.46	8.98	9.04	9.09	11.39
		KH	6.67	8.62	6.00	8.69	8.72	8.67	11.26
		KO	5.90	8.28	5.67	8.31	8.38	8.70	10.68
		RT	5.30	7.67	6.03	8.12	8.30	8.03	10.62
		KE	11.07	9.72	7.07	9.72	9.92	9.94	12.23

			V room						
Measure	***Content***	***A room***	**off**	**GH**	**JC**	**KH**	**KO**	**RT**	**KE**
STD	Music (n = 50)	off	—	4.36	3.04	3.53	2.59	3.54	3.61
		GH	3.54	3.77	2.40	3.03	3.07	3.10	3.83
		JC	4.03	3.78	2.74	3.11	3.01	2.94	3.63
		KH	4.19	3.15	2.90	2.83	3.20	3.05	3.69
		KO	3.09	4.24	2.69	2.77	3.03	2.88	4.42
		RT	2.93	3.08	2.73	3.15	3.39	3.30	4.04
		KE	4.83	4.12	3.90	3.57	3.58	4.17	3.72
	Speech (n = 38)	off	—	4.10	2.02	2.76	4.25	3.90	4.10
		GH	3.52	3.15	2.28	2.74	2.67	2.95	4.44
		JC	4.17	3.01	2.55	2.81	2.50	3.02	3.95
		KH	3.04	3.46	2.12	2.73	2.36	2.44	4.48
		KO	2.75	2.81	1.93	2.63	2.57	3.08	4.54
		RT	3.34	3.21	2.55	2.38	2.91	2.67	4.83
		KE	5.82	3.70	2.73	3.01	3.37	3.60	4.03

Table 11.
Descriptive statistics of perceived source distance *($\hat{D}$).*

			V room						
Measure	***Content***	***A room***	**off**	**GH**	**JC**	**KH**	**KO**	**RT**	**KE**
Mean	Music (n = 50)	off	—	30.76	24.39	18.36	25.03	18.79	32.28
		GH	16.89	29.38	22.66	19.48	22.94	19.32	30.06
		JC	18.36	28.97	24.10	19.05	23.75	19.34	30.63
		KH	14.99	29.95	22.94	18.96	22.22	18.97	29.94
		KO	12.51	28.70	22.78	18.20	22.07	18.88	28.26
		RT	10.56	27.02	22.66	17.99	22.45	18.11	28.13
		KE	31.88	31.52	26.18	21.73	25.64	23.12	33.40
	Speech (n = 38)	off	—	29.29	21.56	17.18	21.41	17.43	32.04
		GH	16.19	26.51	22.53	17.71	20.37	18.69	29.28
		JC	20.75	27.78	23.43	19.07	20.69	19.34	30.03
		KH	11.53	25.70	21.71	17.15	19.36	17.39	28.98
		KO	9.85	25.07	20.89	16.29	19.54	17.00	26.69
		RT	8.42	24.60	20.91	17.04	18.94	16.96	26.69
		KE	31.93	29.47	26.04	20.85	23.25	21.30	33.11
STD	Music (n = 50)	off	—	7.01	7.09	6.67	7.45	6.79	8.76
		GH	6.58	8.39	6.21	5.62	7.40	7.02	9.12
		JC	7.05	8.89	7.96	6.50	7.10	6.66	8.73
		KH	6.30	8.06	7.10	6.36	7.17	5.69	8.73
		KO	6.36	7.88	7.02	5.49	6.74	6.57	9.00
		RT	4.83	9.12	7.24	5.54	7.51	6.54	8.96
		KE	8.67	7.91	7.91	8.58	8.21	8.68	9.20

Measure	*Content*	*A room*	*V room*						
			off	**GH**	**JC**	**KH**	**KO**	**RT**	**KE**
	Speech (n = 38)	off	—	9.68	7.31	5.97	9.02	7.43	9.43
		GH	7.28	10.02	7.79	8.06	8.25	8.15	11.07
		JC	7.56	9.79	8.60	7.53	8.09	8.02	11.42
		KH	5.94	11.32	10.34	7.93	7.68	7.59	12.80
		KO	3.48	11.35	9.37	6.92	8.40	7.91	12.47
		RT	5.89	10.99	10.12	8.07	9.18	8.03	12.94
		KE	8.39	9.36	8.82	10.39	9.82	9.66	10.71

Table 12.
Descriptive statistics of perceived room size *($\hat{S}$).*

Author details

Hans-Joachim Maempel* and Michael Horn
Federal Institute for Music Research, Berlin, Germany

*Address all correspondence to: maempel@sim.spk-berlin.de

References

[1] Cabrera D, Nguyen A, Choi YJ. Auditory versus visual spatial impression: A study of two auditoria. In: Barrass S, Vickers P, editors. Proc. of ICAD 04-Tenth Meeting of the Int. Conf. on Auditory Display. Sydney, Australia: International Community for Auditory Display (ICAD); 2004

[2] Kuusinen A, Lokki T. Auditory distance perception in concert halls and the origins of acoustic intimacy. Proceedings of the Institute of Acoustics. 2015;**37**(3):151-158

[3] Hyde JA. Discussion of the relation between initial time delay gap (ITDG) and acoustical intimacy: Leo Beranek's final thoughts on the subject, documented. Acoustics. 2019;**1**(3): 561-569. DOI: 10.3390/acoustics1030032

[4] Stevens JC, Marks LE. Cross-modality matching of brightness and loudness. Proceedings of the National Academy of Sciences of the USA. 1965;**2**: 407-411

[5] Thomas GJ. Experimental study of the influence of vision on sound localization. Journal of Experimental Psychology. 1941;**28**:163-177. DOI: 10.1037/h0055183

[6] Howard IP, Templeton WB. Human Spatial Orientation. London: Wiley; 1966

[7] Gardner MB. Proximity image effect in sound localization. The Journal of the Acoustical Society of America. 1968; **43**(1):163. DOI: 10.1121/1.1910747

[8] Mateeff S, Hohnsbein J, Noack T. Dynamic visual capture: Apparent auditory motion induced by a moving visual target. Perception. 1985;**14**: 721-727. DOI: 10.1068/p140721

[9] Kitajima N, Yamashita Y. Dynamic capture of sound motion in three-dimensional space. Perceptual and Motor Skills. 1999;**89**(3):1139-1158. DOI: 10.2466/pms.1999.89.3f.1139

[10] Kohlrausch A, van de Par S. Audio-visual interaction in the context of multi-media applications. In: Blauert J, editor. Communication Acoustics (Chapter 5). Berlin: Springer; 2005. pp. 109-138. DOI: 10.1007/3-540-27437-5_5

[11] Shams L, Kamitani Y, Shimojo S. Visual illusion induced by sound. Cognitive Brain Research. 2002;**14**: 147-152. DOI: 10.1016/S0926-6410(02) 00069-1

[12] Andersen TS, Tippana K, Sams M. Factors influencing audiovisual fission and fusion illusions. Cognitive Brain Research. 2004;**21**:301-308. DOI: 10.1016/j.cogbrainres.2004. 06.004

[13] Vatakis A, Spence C. Audiovisual synchrony perception for music, speech, and object actions. Brain Research. 2006;**1111**(1):134-142. DOI: 10.1016/j. brainres.2006.05.078

[14] MacDonald J, McGurk H. Visual influences on speech perception process. Perception & Psychophysics. 1978;**24**:253-257. DOI: 10.3758/ BF03206096

[15] Beerends JG, de Caluwe FE. The influence of video quality on perceived audio quality and vice versa. Journal of the Audio Engineering Society. 1999;**47**: 355-362

[16] Larsson P, Västfjäll D, Kleiner M. Auditory-visual interaction in real and virtual rooms. In: 3rd Convention of the EAA. Spain: Sevilla; 2002

[17] Larsson P, Väljamäe A. Auditory-visual perception of room size in virtual environments. In: Proc. of the 19th Int.

Congress on Acoustics. Madrid; 2007 PPA-03-001

[18] Maempel H-J, Jentsch M. Auditory and visual contribution to egocentric distance and room size perception. Building Acoustics. 2013;**20**(4):383-401. DOI: 10.1260/1351-010X.20.4.383

[19] Treisman A. The binding problem. Current Opinion in Neurobiology. 1996; **6**(2):171-178. DOI: 10.1016/S0959-4388(96)80070-5

[20] Bishop ID, Rohrmann B. Subjective responses to simulated and real environments: A comparison. Landscape and Urban Planning. 2003; **65**(4):261-277. DOI: 10.1016/S0169-2046(03)00070-7

[21] de Kort YAW, IJsselsteijn WA, Kooijman J, Schuurmans Y. Virtual laboratories: Comparability of real and virtual environments for environmental psychology. Presence—Teleoperators and Virtual Environments. 2003;**12**(4): 360-373. DOI: 10.1162/105474603322391604

[22] Billger M, Heldal I, Stahre B, Renstrom K. Perception of color and space in virtual reality: a comparison between a real room and virtual reality models. In: Rogowitz BE, Pappas TN, editors. Proc. of SPIE, Human Vision and Electronic Imaging IX, San Jose, California, USA. Vol. 5292. Bellingham, WA: Society of Photographic Instrumentation Engineers (SPIE); 2004. pp. 90-98. DOI: 10.1117/12.526986

[23] Kuliga SF, Thrash T, Dalton RC, Hölscher C. Virtual reality as an empirical research tool – Exploring user experience in a real building and a corresponding virtual model. Computers, Environment and Urban Systems. 2015;**54**:363-375. DOI: 10.1016/j.compenvurbsys.2015.09.006

[24] Nielsen SH. Auditory distance perception in different rooms. Journal of the Audio Engineering Society. 1993;**41**: 755-770

[25] Bronkhorst AW, Houtgast T. Auditory distance perception in rooms. Nature. 1999;**397**:517-520. DOI: 10.1038/17374

[26] Bronkhorst AW, Zahorik P. The direct-to-reverberant ratio as cue for distance perception in rooms. The Journal of the Acoustical Society of America. 2002;**111**(5):2440-2441. DOI: 10.1121/1.4809156

[27] Loomis JM, Klatzky RL, Golledge RG. Auditory distance perception in real, virtual, and mixed environments. In: Ohta Y, Tamura H, editors. Mixed Reality: Merging Real and Virtual Worlds. Ohmsha: Tokyo; 1999. pp. 201-214

[28] Zahorik P, Brungart DS, Bronkhorst AW. Auditory distance perception in humans: A summary of past and present research. Acta Acustica united with Acustica. 2005;**91**(3): 409-420

[29] Kolarik AJ, Cirstea S, Pardhan S. Discrimination of virtual auditory distance using level and direct-to-reverberant ratio cues. The Journal of the Acoustical Society of America. 2013; **134**(5):3395. DOI: 10.1121/1.4824395

[30] Kolarik AJ, Moore BCJ, Zahorik P, Cirstea S, Pardhan S. Auditory distance perception in humans: A review of cues, development, neuronal bases, and effects of sensory loss. Attention, Perception, & Psychophysics. 2016;**78**: 373-395. DOI: 10.3758/s13414-015-1015-1

[31] Moulin S, Nicol R, Gros L. Auditory distance perception in real and virtual environments. In: SAP '13, Proc. of the ACM Symposium on Applied Perception. Dublin: ACM; 2013. p. 117. DOI: 10.1145/2492494.2501876

[32] Kearney G, Gorzel M, Boland F, Rice H. Depth perception in interactive virtual acoustic environments using higher order ambisonic sound fields. In: 2nd Int. Symposium on Ambisonics and Spherical Acoustics. Berlin: Univ.-Verl. TU; 2010

[33] Calcagno ER, Abregu EL, Eguia MC, Vergara R. The role of vision in auditory distance perception. Perception. 2012; **41**(2):175-192. DOI: 10.1068/p7153

[34] Zahorik P. Assessing auditory distance perception using virtual acoustics. The Journal of the Acoustical Society of America. 2002;**111**(4): 1832-1846. DOI: 10.1121/1.1458027

[35] Chan JS, Lisiecka D, Ennis C, O'Sullivan C, Newell FN. Comparing audiovisual distance perception in various 'real' and 'virtual' environments. Perception ECVP Abstract. 2009;**38**:30

[36] Rébillat M, Boutillon X, Corteel È, Katz BFG. Audio, visual, and audio-visual egocentric distance perception in virtual environments. In: EAA Forum Acusticum 2011. Denmark: Aalborg; 2011. pp. 482-487

[37] Rébillat M, Boutillon X, Corteel È, Katz BFG. Audio, visual, and audio-visual egocentric distance perception by moving subjects in virtual environments. ACM Transactions on Applied Perception. 2012;**9**:19:1-19:17. DOI: 10.1145/2355598.2355602

[38] Maempel H-J, Horn M. Audiovisual perception of real and virtual rooms. Journal of Virtual Reality and Broadcasting. 2017;**14**(5):1-15. DOI: 10.20385/1860-2037/14.2017.5

[39] Cabrera D, Jeong D, Kwak HJ, Kim J-Y. Auditory room size perception for modelled and measured rooms. In: Internoise, the 2005 Congress and Exposition on Noise Control Engineering. Rio de Janeiro, Brazil: Institute of Noise Control Engineering - USA (INCE-USA); 2005

[40] Cabrera D, Pop C, Jeong D. Auditory room size perception: a comparison of real versus binaural sound-fields. In: Acoustics. Christchurch, New Zealand; 2006. pp. 417-422

[41] Cabrera D. Acoustic clarity and auditory room size perception. In: 14th Int. Congress on Sound & Vibration. Cairns, Australia; 2007

[42] Hameed S, Pakarinen J, Valde K, Pulkki V. Psychoacoustic cues in room size perception. In: AES 116th Convention. Berlin, Germany: Audio Engineering Society; 2004. Convention Paper 6084

[43] Yadav M, Cabrera D, Martens WL. Auditory room size perceived from a room acoustic simulation with autophonic stimuli. Acoustics Australia. 2011;**39**(3):101-105

[44] Cutting JE, Vishton PM. Perceiving layout and knowing distances: the integration, relative potency, and contextual use of different information about depth. In: Epstein W, Rogers S, editors. Perception of Space and Motion (Chapter 3). San Diego et al: Academic Press; 1995. pp. 69-117. DOI: 10.1016/B978-012240530-3/50005-5

[45] Mehrabi M, Peek EM, Wuensche BC, Lutteroth C. Making 3D work: A classification of visual depth cues, 3D display technologies and their applications. Proc. of the 14th Australasian User Interface conference (AUIC 2013), Adelaide, Australia. Conferences in Research and Practice in Information Technology (CRPIT). 2013; **139**:91-100. DOI: 10.5555/2525493.2525503

[46] Renner RS, Velichkovsky BM, Helmert JR. The perception of egocentric distances in virtual environments—A review. ACM

Computing Surveys. 2013;**46**(2):1-40. DOI: 10.1145/2543581.2543590

[47] Zahorik P. Audio/visual interaction in the perception of sound source distance. In: Ochmann M, Vorländer M, Fels J, editors. ICA 2019 Aachen. Proc. of the 23rd Int. Congress on Acoustics, Aachen, Germany. Berlin: Deutsche Gesellschaft für Akustik; 2019. pp. 7927-7931

[48] Loomis JM, Da Silva JA, Philbeck JW, Fukusima SS. Visual perception of location and distance. Current Directions in Psychological Science. 1996;**5**(3):72-77. DOI: 10.1111/1467-8721.ep10772783

[49] Loomis JM, Knapp JM. Visual perception of egocentric distance in real and virtual environments. In: Hettinger LJ, Haas MW, editors. Virtual and Adaptive Environments. Applications, Implications, and Human Performance Issues. Mahwah, NJ: Erlbaum; 2003. pp. 21-46

[50] Plumert JM, Kearney JK, Cremer JF, Recker K. Distance perception in real and virtual environments. ACM Transactions on Applied Perception. 2005;**2**(3):216-233. DOI: 10.1145/1077399.1077402

[51] Armbrüster C, Wolter M, Kuhlen T, Spijkers W, Fimm B. Depth perception in virtual reality: Distance estimations in peri- and extrapersonal space. CyberPsychology & Behavior. 2008; **11**(1):9-15. DOI: 10.1089/cpb.2007.9935

[52] Klein E, Swan JE, Schmidt GS, Livingston MA, Staadt OG. Measurement protocols for medium-field distance perception in large-screen immersive displays. In: IEEE Virtual Reality. Vol. 2009. Lafayette, Lousiana, USA; 2009. pp. 107-113. DOI: 10.1109/VR.2009.4811007

[53] Naceri A, Chellali R, Dionnet F, Toma S. Depth perception within virtual environments: A comparative study between wide screen stereoscopic displays and head mounted devices. In: 2009 Computation World: Future Computing, Service Computation, Cognitive, Adaptive, Content, Patterns. 2009. pp. 460-466. DOI: 10.1109/ComputationWorld.2009.91

[54] Ziemer CJ, Plumert JM, Cremer JF, Kearney JK. Estimating distance in real and virtual environments: Does order make a difference? Attention, Perception, & Psychophysics. 2009;**71**(5):1095-1106. DOI: 10.3758/APP.71.5.1096

[55] Alexandrova IV, Teneva PT, de la Rosa S, Kloos U, Bülthoff HH, Mohler BJ. Egocentric distance judgments in a large screen display immersive virtual environment. In: 7th Symposium on Applied Perception in Graphics and Visualization. Los Angeles, CA, USA; 2010. pp. 57-60. DOI: 10.1145/1836248.1836258

[56] Interrante V, Anderson L, Ries B. Distance perception in immersive virtual environments, revisited. In: Proc. of the IEEE Virtual Reality Conf. (VR'06). New York: IEEE; 2006. pp. 3-10. DOI: 10.1109/VR.2006.52

[57] Bruder G, Argelaguet F, Olivier AH, Lécuyer A. CAVE size matters: Effects of screen distance and parallax on distance estimation in large immersive display setups. Presence—Teleoperators and Virtual Environments. 2016;**25**(1): 1-16. DOI: 10.1162/PRES_a_00241

[58] Gadia D, Galmonte A, Agostini T, Viale A, Marini D. Depth and distance perception in a curved, large screen virtual reality installation. In: Woods AJ, Holliman NS, Merritt JO, editors. Stereoscopic Displays and Applications XX. Proc. of SPIE. 2009;**7237**:723718. DOI: 10.1117/12.805809

[59] Creem-Regehr SH, Willemsen P, Gooch AA, Thompson WB. The

influence of restricted viewing conditions on egocentric distance perception: Implications for real and virtual indoor environments. Perception. 2005;**34**(2):191-204. DOI: 10.1068/p5144

[60] Kruszielski LF, Kamekawa T, Marui A. The influence of camera focal length in the direct-to-reverb ratio suitability and its effect in the perception of distance for a motion picture. In: AES 131st Convention. New York; 2011. Convention paper 8580.

[61] Anderson PW, Zahorik P. Auditory/visual distance estimation. Accuracy and variability. Front Psychol. 2014;**5**: 1097. DOI: 10.3389/fpsyg.2014.01097

[62] Larsson P, Västfjäll D, Kleiner M. Ecological acoustics the multimodal perception of rooms – real and unreal experiences of auditory-visual virtual environments. In: Hiipakka J, Zacharov N, Takala T, editors. 2001 Int. Conf. on Auditory Display. Espoo, Finland: Helsinki University of Technology; 2001

[63] Thurlow WR, Jack CE. Certain determinants of the "ventriloquism effect". Perceptual and Motor Skills. 1973;**36**(suppl. 3):1171-1184. DOI: 10.2466/pms.1973.36.3c.1171

[64] Jack CE, Thurlow WR. Effects of degree of visual association and angle of displacement on the "ventriloquism" effect. Perceptual and Motor Skills. 1973;**37**(3):967-979. DOI: 10.1177/003151257303700360

[65] Chen L, Vroomen J. Intersensory binding across space and time. A tutorial review. Attention, Perception, & Psychophysics. 2013;**75**(5):790-811. DOI: 10.3758/s13414-013-0475-4

[66] Mershon DH, Desaulniers DH, Amerson TLJ, Kiefer SA. Visual capture in auditory distance perception. Proximity image effect reconsidered. The Journal of Auditory Research. 1980; **20**:129-136

[67] Zahorik P. Estimating sound source distance with and without vision. Optometry and Vision Science. 2001; **78**(5):270-275

[68] Côté N, Koehl V, Paquier M. Ventriloquism effect on distance auditory cues. In: Acoustics 2012 Joint Congress (11ème Congrès Français d'Acoustique—2012 Annual IOA Meeting), Apr 2012. France: Nantes; 2012. pp. 1063-1067

[69] Postma BNJ, Katz BFG. The influence of visual distance on the room-acoustic experience of auralizations. The Journal of the Acoustical Society of America. 2017; **142**(5):3035-3046. DOI: 10.1121/1.5009554

[70] Ernst MO, Banks MS. Humans integrate visual and haptic information in a statistically optimal fashion. Nature. 2002;**415**:429-433. DOI: 10.1038/415429a

[71] Battaglia PW, Jacobs RA, Aslin RN. Bayesian integration of visual and auditory signals for spatial localization. Journal of the Optical Society of America. A. 2003;**20**(7):1391-1397. DOI: 10.1364/josaa.20.001391

[72] Alais D, Burr D. The ventriloquist effect results from near-optimal bimodal integration. Current Biology. 2004;**14**(3):257-262. DOI: 10.1016/j.cub.2004.01.029

[73] Finnegan DJ, Proulx MJ, O'Neill E. Compensating for distance compression in audiovisual virtual environments using incongruence. In: Proc. of the 2016 CHI Conf. on Human Factors in Computing Systems (CHI'16). New York: ACM; 2016. pp. 200-212. DOI: 10.1145/2858036.2858065

[74] Agganis BT, Muday JA, Schirillo JA. Visual biasing of auditory localization in azimuth and depth. Perceptual and Motor

Skills. 2010;**111**(3):872-892. DOI: 10.2466/22.24.27.PMS.111.6.872-892

[75] Hládek L, Le Dantec CC, Kopčo N, Seitz A. Ventriloquism effect and aftereffect in the distance dimension. Proceedings of Meetings on Acoustics. 2013;**19**:050042. DOI: 10.1121/1.4799881

[76] Roach NW, Heron J, McGraw PV. Resolving multisensory conflict: A strategy for balancing the costs and benefits of audio-visual integration. Proceedings of the Royal Society of London. Series B. 2006;**273**(1598):2159-2168. DOI: 10.1098/rspb.2006.3578

[77] Meijer D, Veselič S, Calafiore C, Noppeney U. Integration of audiovisual spatial signals is not consistent with maximum likelihood estimation. Cortex. 2019;**119**:74-88. DOI: 10.1016/j.cortex.2019.03.026

[78] André CR, Corteel E, Embrechts JJ, Verly JG, Katz BFG. Subjective evaluation of the audiovisual spatial congruence in the case of stereoscopic-3D video and wave field synthesis. International Journal of Human-Computer Studies. 2014;**72**(1):23-32. DOI: 10.1016/j.ijhcs.2013.09.004

[79] Mendonça C, Mandelli P, Pulkki V. Modelling the perception of audiovisual distance. Bayesian causal inference and other models. PLoS One. 2016;**11**(12):e0165391. DOI: 10.1371/journal.pone.0165391

[80] Spence C. Crossmodal correspondences: A tutorial review. Attention, Perception, & Psychophysics. 2011;**73**:971-995. DOI: 10.3758/s13414-010-0073-7

[81] Maempel H-J. Apples and oranges: A methodological framework for basic research into audiovisual perception. In: Hohmaier S, editor. Jahrbuch des Staatl. Inst. für Musikforschung 2016. Mainz et al: Schott; 2019. pp. 361–377. DOI: 10.14279/depositonce-6424.2

[82] Berg J, Rumsey F. Verification and correlation of attributes used for describing the spatial quality of reproduced sound. In: AES 19th Int. Conf., Schloss Elmau, Germany. Article No. 1932. New York City, New York, USA: Audio Engineering Society; 2001

[83] Rumsey F. Spatial quality evaluation for reproduced sound: Terminology, meaning, and a scene-based paradigm. Journal of the Audio Engineering Society. 2002;**50**(9):651-666

[84] Bizley JK, Maddox RK, Lee AKC. Defining auditory-visual objects: Behavioral tests and physiological mechanisms. Trends in Neurosciences. 2016;**39**(2):74-85. DOI: 10.1016/j.tins.2015.12.007

[85] Faul F, Erdfelder E, Lang A-G, Buchner A. G*power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behavior Research Methods. 2007;**39**:175-191. DOI: 10.3758/BF03193146

[86] Rasch B, Friese M, Hofmann W, Naumann E. Quantitative Methoden. 3rd ed. Vol. 2. Heidelberg: Springer; 2010

[87] Deutsches Institut für Normung e.V. DIN EN ISO 3382-1 Akustik – Messung von Parametern der Raumakustik – Teil 1: Aufführungsräume. Berlin: Beuth; 2009

[88] Beranek L. Concert hall acoustics. Journal of the Audio Engineering Society. 2008;**56**(7/8):532-544

[89] Harris CM. Absorption of sound in air versus humidity and temperature. The Journal of the Acoustical Society of America. 1966;**40**(1):148-159. DOI: 10.1121/1.1910031

[90] Lindau A, Schultz F, Horn M, Brinkmann F, Erbes V, Fuß A, Maempel H-J, Weinzierl S. Raumakustische Messungen in sechs Aufführungsräumen: Konzerthaus/ Kleiner Saal (Berlin), Jesus-Christus-Kirche (Berlin), Kloster Eberbach/ Basilika (Eltville am Rhein), Renaissance-Theater (Berlin), Komische Oper (Berlin), Gewandhaus/Großer Saal (Leipzig), measurement reports (research data). 2021:i-VI/16. DOI: 10.14279/depositonce-11947

[91] Lindau A, Weinzierl S. FABIAN – An instrument for software-based measurement of binaural room impulse responses in multiple degrees of freedom. In: 24. Leipzig: Tonmeistertagung; 2006

[92] Lindau A, Hohn T, Weinzierl S. Binaural resynthesis for comparative studies of acoustical environments. In: AES 122nd Convention. Vienna; 2007. Preprint 7032

[93] Hendrickx E, Stitt P, Messonier J-C, Lyzwa J-M, Katz BFG, de Boishéraud C. Influence of head tracking on the externalization of speech stimuli for non-individualized binaural synthesis. The Journal of the Acoustical Society of America. 2017;**141**(3):2011-2023. DOI: 10.1121/1.4978612

[94] McAnally KI, Martin RL. Sound localization with head movement: Implications for 3-d audio displays. Frontiers in Neuroscience. 2014;**8**(210): 1-6. DOI: 10.3389/fnins.2014.00210

[95] Lindau A, Maempel H-J, Weinzierl S. Minimum BRIR grid resolution for dynamic binaural synthesis. In: Forum Acusticum, European Acoustics Association. Proc. of Acoustics'08, Conference: Paris. Stuttgart: Hirzel; 2008. pp. 3851-3856

[96] Schultz F, Lindau A, Weinzierl S. Just noticeable BRIR grid resolution for lateral head movements. In: DAGA 2009. Rotterdam; 2009. pp. 200-201

[97] Erbes V, Schultz F, Lindau A, Weinzierl S. An extraaural headphone system for optimized binaural reproduction. In: DAGA 2012, Darmstadt. Berlin: Deutsche Gesellschaft für Akustik; 2012. pp. 313-314

[98] Lindau A, Brinkmann F. Perceptual evaluation of headphone compensation in binaural synthesis based on non-individual recordings. Journal of the Audio Engineering Society. 2012;**60**(1/2):54-62

[99] Lindau A. The perception of system latency in dynamic binaural synthesis. In: DAGA 2009, Rotterdam. Berlin: Deutsche Gesellschaft für Akustik; 2009. pp. 1063-1066

[100] Lindau A, Estrella J, Weinzierl S. Individualization of dynamic binaural synthesis by real time manipulation of the ITD. In: AES 128th Convention. London: Audio Engineering Society; 2010. Preprint 8088

[101] Neidhardt A, Reif B. Minimum BRIR grid resolution for interactive position changes in dynamic binaural synthesis. In: Proc. of the 148th Int. AES Convention, Vienna, Austria. New York City, New York, USA: Audio Engineering Society; 2020. Available from: https://www.aes.org/e-lib/browse.cfm?elib=20788

[102] Werner S, Klein F, Neidhardt A, Sloma U, Schneiderwind C, Brandenburg K. Creation of auditory augmented reality using a position-dynamic binaural synthesis system—Technical components, psychoacoustic needs, and perceptual evaluation. Applied Sciences. 2021;**11**(3):1150. DOI: 10.3390/app11031150

[103] Speigle JM, Loomis JM. Auditory distance perception by translating

observers. In: Proc. of 1993 IEEE Research Properties in Virtual Reality Symposium. New York: IEEE; 1993. pp. 92-99. DOI: 10.1109/VRAIS.1993.378257

[104] Genzel D, Schutte M, Brimijoin WO, MacNeilage PR, Wiegrebe L. Psychophysical evidence for auditory motion parallax. PNAS. 2018;**115**(16):4264-4269. DOI: 10.1073/pnas.1712058115

[105] Carlile S, Leung J. The perception of auditory motion. Trends in Hearing. 2016;**20**:1-19. DOI: 10.1177/2331216516644254

[106] Rummukainen OS, Schlecht SJ, Habets EAP. Self-translation induced minimum audible angle. The Journal of the Acoustical Society of America. 2018; **144**(4):EL340–EL345. DOI: 10.1121/1.5064957

[107] Rosenberg LB. The effect of interocular distance upon operator performance using stereoscopic displays to perform virtual depth tasks. In: Proceedings of IEEE Virtual Reality Annual International Symposium. New York: IEEE; 1993. pp. 27-32. DOI: 10.1109/VRAIS.1993.380802

[108] Utsumi A, Milgram P, Takemura H, Kishino F. Investigation of errors in perception of stereoscopically presented virtual object locations in real display space. Proceedings of the Human Factors and Ergonomics Society Annual Meeting. 1994;**38**(4):250-254. DOI: 10.1177/154193129403800413

[109] Best S. Perceptual and oculomotor implications of interpupillary distance settings on a head-mounted virtual display. Proceedings of Naecon IEEE Nat. 1996;**1**:429-434. DOI: 10.1109/NAECON.1996.517685

[110] Drascic D, Milgram P. Perceptual issues in augmented reality. In: Proc. SPIE 2653, Stereoscopic Displays and Virtual Reality Systems III. 1996. pp. 123-134. DOI: 10.1117/12.237425

[111] Wartell Z, Hodges LF, Ribarsky W. Balancing fusion, image depth and distortion in stereoscopic head-tracked displays. In: SIGGRAPH '99: Proc. of the 26th annual conference on Computer graphics and interactive techniques. 1999. pp. 351-358. DOI: 10.1145/311535.311587

[112] Renner RS, Steindecker E, Müller M, Velichkovsky BM, Stelzer R, Pannasch S, et al. The influence of the stereo base on blind and sighted reaches in a virtual environment. ACM Transactions on Applied Perception. 2015;**12**(2):1-18. DOI: 10.1145/2724716

[113] Kim N-G. Independence of size and distance in binocular vision. Frontiers in Psychology. 2018;**9**:1-18. DOI: 10.3389/fpsyg.2018.00988

[114] Willemsen P, Gooch AA, Thompson WB, Creem-Regehr SH. Effects of stereo viewing conditions on distance perception in virtual environments. Presence—Teleoperators and Virtual Environments. 2008;**17**(1): 91-101. DOI: 10.1162/pres.17.1.91

[115] Bruder G, Pusch A, Steinicke F. Analyzing effects of geometric rendering parameters on size and distance estimation in on-axis stereographics. In: SAP '12: Proc. of the ACM Symposium on Applied Perception. 2012. pp. 111-118. DOI: 10.1145/2338676.2338699

[116] Deutsches Institut für Normung e. V. DIN 33402–2 Ergonomie – Körpermaße Des Menschen – Teil 2: Werte. Berlin: Beuth; 2020

[117] Knapp JM, Loomis JM. Limited field of view of head-mounted displays is not the cause of distance underestimation in virtual environments. Presence—Teleoperators and Virtual Environments. 2004;**13**(5):

572-577. DOI: 10.1162/105474604254 5238

[118] Maempel HJ, Horn M. The virtual concert hall – A research tool for the experimental investigation of audiovisual room perception. International Journal on Stereoscopic and Immersive Media. 2017;**1**(1):78-98

[119] Lindau A, Weinzierl S. Assessing the plausibility of virtual acoustic environments. Acta Acustica United with Acustica. 2012;**98**(5):804-810. DOI: 10.3813/AAA.918562

[120] Olejnik S, Algina J. Generalized eta and omega squared statistics: Measures of effect size for some common research designs. Psychological Methods. 2003; **8**(4):434-447. DOI: 10.1037/1082-989X. 8.4.434

[121] Bakeman R. Recommended effect size statistics for repeated measures designs. Behavior Research Methods. 2005;**37**(3):379-384. DOI: 10.3758/ BF03192707

[122] Cohen J. Statistical Power Analysis for the Behavioral Sciences. 2nd ed. New Jersey: Erlbaum; 1988

[123] Lakens D. Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for *t*-tests and ANOVAs. Frontiers in Psychology. 2013;**4**(863):1-12. DOI: 10.3389/fpsyg.2013.00863

[124] Cabrera D. Control of perceived room size using simple binaural technology. In: Martens WL, editor. Proc. of the 13th Int. Conf. on Auditory Display. Montréal, Canada: International Community for Auditory Display; 2007

[125] FAQ: Humanities and Social Sciences. Statement by an Ethics Committee [Internet]. Available from: https://www.dfg.de/en/research_fund ing/faq/faq_humanities_social_science/ index.html [Accessed: January 10, 2022]

[126] Berufsethische Richtlinien [Internet]. In: Berufsverband Deutscher Psychologinnen und Psychologen e.V. Deutsche Gesellschaft für Psychologie e.V. 2016. Available from: https://www. dgps.de/fileadmin/user_upload/PDF/ berufsethik-foederation-2016.pdf [Accessed: January 10, 2022]