The percept of reverberation is not affected by visual room impression in virtual environments

Michael Schutte, Stephan D. Ewert, and Lutz Wiegrebe

Citation: The Journal of the Acoustical Society of America 145, EL229 (2019); doi: 10.1121/1.5093642
View online: https://doi.org/10.1121/1.5093642
View Table of Contents: https://asa.scitation.org/toc/jas/145/3
Published by the Acoustical Society of America

ARTICLES YOU MAY BE INTERESTED IN

Spectral manipulation improves elevation perception with non-individualized head-related transfer functions
The Journal of the Acoustical Society of America 145, EL222 (2019); https://doi.org/10.1121/1.5093641

Raking early reflection signals for late reverberation and noise reduction
The Journal of the Acoustical Society of America 145, EL257 (2019); https://doi.org/10.1121/1.5095535

Age effects on the contributions of envelope and periodicity cues to recognition of interrupted speech in quiet and with a competing talker
The Journal of the Acoustical Society of America 145, EL173 (2019); https://doi.org/10.1121/1.5091664

Volumetric reconstruction of acoustic energy flows in a reverberation room
The Journal of the Acoustical Society of America 145, EL203 (2019); https://doi.org/10.1121/1.5092820

Investigation on the diffusive surface modeling detail in geometrical acoustics based simulations
The Journal of the Acoustical Society of America 145, EL215 (2019); https://doi.org/10.1121/1.5092821

A deep learning algorithm to increase intelligibility for hearing-impaired listeners in the presence of a competing talker and reverberation
The Journal of the Acoustical Society of America 145, 1378 (2019); https://doi.org/10.1121/1.5093547
The percept of reverberation is not affected by visual room impression in virtual environments

Michael Schutte,1,a) Stephan D. Ewert,2 and Lutz Wiegrebe1

1Division of Neurobiology, Department Biology II and Graduate School of Systemic Neurosciences, Ludwig-Maximilians-Universität München, Germany
2Medical Physics and Cluster of Excellence Hearing4all, University of Oldenburg, Germany
michael.schutte@uiae.at, stephan.ewert@uni-oldenburg.de, lutzw@lmu.de

Abstract: Humans possess mechanisms to suppress distracting early sound reflections, summarized as the precedence effect. Recent work shows that precedence is affected by visual stimulation. This paper investigates possible effects of visual stimulation on the perception of later reflections, i.e., reverberation. In a highly immersive audio-visual virtual reality environment, subjects were asked to quantify reverberation in conditions where simultaneously presented auditory and visual stimuli either match in room identity, sound source azimuth, and sound source distance, or diverge in one of these aspects. While subjects reliably judged reverberation across acoustic environments, the visual room impression did not affect reverberation estimates.
parameters based on photos (McCreery and Calamia, 2006), suggesting that we possess an intuitive awareness of the acoustics of rooms which we can see, but not hear.

Supporting the notion of multi-modal integration, it was demonstrated that, well beyond the classical ventriloquism effect, the visual system may also affect the way we deal with reflections of a sound source in enclosed spaces. It was shown that the strength of the precedence effect can be enhanced when the layout of a visual environment is consistent with the acoustically presented sounds and their reflections. Likewise, the precedence effect is diminished when visual and auditory environments are inconsistent (Bishop et al., 2011, 2012). The precedence effect is typically relevant for suppressing spatial information of early reflections. However, it is to date unclear whether the visual impression of a room may affect the perception of later reflections which overlap in the reverberant tail of a room response. Depending on the temporal decay of the reverberation, technically characterized by the reverberation time (RT60) and the energetic ratio in relation to the direct sound [direct-to-reverberant ratio (DRR)] and early reflections, human can judge the perceived reverberation time and level [e.g., Lindau et al. (2014)]. Extrapolating from the documented effects of the visual system on classical precedence, we assess here whether perceived reverberation is affected by the congruence of the visual and auditory environments.

To this end, we quantified the extent to which subjects perceive the same auditory environment as less or more reverberant when the visual and auditory environments are congruent in comparison to a condition where the visual environment is not shown, or where it is incongruent with the auditory environment. We recruited latest audio-visual stimulation techniques to create highly realistic and immersive environments, and thus ensure that possible effects may be relevant in every-day listening situations.

### 2. Methods

#### 2.1 Subjects and reproduction setup

Ten listeners (21–27 years of age, mean 24.3, 5 female) participated in the experiment. They were paid for their participation. Procedures were approved by the ethics committee of the Faculty of Medicine, LMU Munich (project No. 18–327).

The listeners were asked to quantify the perceived degree of reverberation in audiovisual and baseline audio-only conditions on a 1–10 integer rating scale. Five repetitions were measured for each trial. The stimulus in each trial paired one of 12 auditory environments with one of 12 visual environments (or the lack of a visual environment), although not all possible combinations were used (see below).

Listeners were seated in an anechoic chamber (2 m × 2 m base, 2.2 m high) and the auditory stimuli were presented via 36 loudspeakers (Plus XS.2, CANTON Elektronik, Weilrod, DE) mounted at head height near the chamber wall in a horizontal circular arrangement at 10° intervals in azimuth. The speakers were driven by four 12-channel power amplifiers (CI9120, NAD Electronics International, Pickering ON, CA) which received input from a PC via two 24-channel audio interfaces (24I/O, MOTU, Cambridge MA, US) running at a sampling rate of 48 kHz. The loudspeakers were fully equalized in spectral magnitude and phase by application of per-speaker compensation impulse responses. The root-mean-square sound pressure in the loudest conditions was 64 dB sound pressure level.

A head-mounted stereo display (Rift DK2, Oculus VR, Menlo Park CA, US) provided the visual stimuli to the subjects. The reference frames of the virtual visual and auditory environments were aligned with each other through careful placement of the infrared tracking camera. The subject’s position in the virtual environment was kept fixed (i.e., the translational component of the head tracking readings was ignored for the real-time updates), and subjects were instructed to rotate their head in the horizontal plane, but not move it translationally, or otherwise rotate it. This was verified by a supervisor from outside the chamber via an infrared camera. Rotational head-tracking data also confirmed that the subjects complied with these instructions.

#### 2.2 Stimuli

Auditory environments were defined by three variables: room identity (bedroom, office room, or factory hall); sound source azimuth (60° left, 0°, or 30° right); and sound source–listener distance (1 or 3 m). We used an improved version of RAZR (Wendt et al., 2016) to simulate the room acoustics for these 18 environments, employing an image-source model for early reflections (up to third order), a scattering module and a feedback delay network for late reverberation. The direct sound and the early reflections were mapped onto 36 channels (corresponding to the locations of the...
loudspeakers in the experimental chamber) by means of vector-based amplitude panning (Pulkki, 1997). The late reverberation was mapped onto twelve channels (three per chamber wall).

Visual environments were defined by the three variables room identity, visual source azimuth, and visual source–listener distance (with the same respective sets of possible values as listed above). The visual environments were rendered from 3-D geometric models to stereo equirectangular panoramic images with the CYCLES engine for BLENDER (Blender Foundation and community, 2018). They depict rooms with the same dimensions and wall materials as the corresponding auditory environments, with a TV set placed at the visual source position.

The visual and auditory stimuli and the experimental chamber are illustrated in Fig. 1.

2.3 Procedure

Two types of trials were presented in the experiment: audiovisual trials and audio-only trials. Each audiovisual trial combined an auditory environment with a visual environment such that

(a) the auditory and the visual room identities, source positions and sound source–listener distances are pairwise identical (congruent condition), or that
(b) the visual room identity differs from the auditory room identity while the other variables match (room identity incongruence), or that
(c) the visual source azimuth differs from the sound source azimuth while the other variables match (azimuth incongruence), or that
(d) the visual source–listener distance differs from the sound source–listener distance while the other variables match (distance incongruence).

Note that audiovisual trials can be incongruent in azimuth by up to 90° (by combining a 60° left auditory environment with a 30° right visual environment, or vice versa).

All auditory environments were also presented in audio-only trials, where the visual stimulus was a uniformly black image.

Trials of both types were presented in a random order. In each trial, a German-language speech signal extracted from a TV news show (no background music or other sounds; loudness-normalized according to EBU R 128) drawn randomly from a pool of 8 signals was convolved with the 36 impulse responses (all of which contained early reflections, and 6 of which also included late reverberation) for the corresponding auditory environment and played back. Simultaneously, a corresponding video clip was shown through the head-mounted display on the virtual TV screen at the visual source location. The video did not show the human speaker. The subjects

![Fig. 1. Stimuli and experimental setup. (a) Estimated impulse responses for each of the three rooms and corresponding room dimensions, broadband reverberation times, and direct-to-reverberant energy ratios (for both the 1 and 3 m sound source distances). (b) Example image of one condition (visual source azimuth = 0°, visual source–listener distance = 3 m) for the same condition. (c) Layout of the speakers inside the experimental chamber. In each trial, exactly one of the dark-shaded speakers played the direct sound from the virtual sound source, and the twelve speakers with a black band emitted late reverberation. All 36 speakers (mounted at 0° elevation) potentially rendered early reflections.](image-url)
were instructed to look toward the TV screen during stimulus presentation, and to reproduce the up-down orientation of an arrow displayed on the virtual TV screen with the joystick (while the arrow was visible) to ensure that their eyes were open. Subsequently, they were asked to judge the perceived degree of reverberation (“wahrgenommene Verhalltheit”) on a purely numeric 1–10 rating scale.

Prior to the experiment, the subjects were given the opportunity to listen to example stimuli representing each value on the rating scale. No visual stimulation took place during this familiarization phase. The familiarization stimuli were the same speech sounds, played through six speakers at 60° offsets in azimuth. The dry speech signal was convolved with random noise impulse responses with an exponentially decaying envelope for each speaker. The impulse responses ranged in broadband reverberation times similarly to those of the synthetic rooms \(RT_{60} = 100 \text{ to } 4000\) ms. The carrier noise was shaped to be spectrally identical to the average magnitude spectrum of the synthetic room impulse responses.

3. Results

To analyze the results from the audio-only control and audiovisually congruent conditions, we fit a linear model with fixed and random effects (mixed-effects model) to the data, with the rating on the 1–10 scale as the dependent variable. The four independent variables auditory room identity, sound source distance, sound source azimuth, and presence of congruent visual stimulation (“visuals on/off”) were included as fixed effects. First-order interactions of these fixed effects were also modelled. The data were grouped by the random factors subject (allowing the slopes for all fixed effects as well as the intercept to vary) and speech signal (allowing only the intercept to vary). All independent variables were treated as categorical.

An adapted F-test, using the approximation of Satterthwaite (1946) for degrees of freedom, revealed a significant \((z = 0.05)\) main effect of room identity \([F(2, 9) = 246.93, p < 10^{-7}]\), and no other significant main effects (azimuth: \(p = 0.93\), distance: \(p = 0.43\), visuals on/off: \(p = 0.45\)). Per-subject mean ratings and standard deviations are shown in Fig. 2, panels #1–10 (audio-only control conditions in light grey, congruent conditions in dark grey). Estimated marginal mean differences were 3.20 ± 0.22 (s.e.m.) for bedroom-office, and 2.38 ± 0.26 for office-factory. Two-sided t-tests, again using the Satterthwaite approximation, established all pairwise room identity differences as significant (bedroom-office: \(t = 21.08, p < 10^{-6}\); office-factory: \(t = 9.172, p < 10^{-4}\); \(p\)-values Tukey-adjusted for multiple comparisons).

Further analysis also showed a significant interaction between sound source distance and room identity \([F(2, 3217) = 18.00, p < 10^{-7}]\), and no other significant interactions (room identity and azimuth: \(p = 0.87\); room identity and visuals on/off: \(p = 0.07\); azimuth and sound source distance: \(p = 0.14\); azimuth and visuals on/off:

![Fig. 2. (1–10) Individual subjects’ reverberation ratings for the auditory room identities bedroom, office and factory. Light grey bars: mean scores per room and subject in the audio-only control conditions. Dark grey bars: mean scores per room and subject in the audiovisually congruent conditions. Black lines: Standard deviations. (Bottom right) Striped bars: Estimated marginal mean ratings (based on a mixed-effects model with subject and speech signal as random effects) for the three rooms, across all subjects and both audio-only control and audiovisually congruent conditions. Dots: Means conditioned on sound source distance (1 m vs 3 m). Black lines: 95% confidence intervals. Asterisks indicate significantly different ratings at \(z = 0.05\).]
this is unlikely to have taken place considering the magnitude of the azimuthal conditions where the visual source position deviates from the sound source position, employed in this experiment is theoretically suitable to elicit a visual capture effect in as mediated by the auditory stimulation. While the audio-visual stimulation technique ated by the visual stimulation (the position of the TV set in the virtual room), but also our subjects to determine the azimuthal position of the sound source not only as medi-

simultaneous visual impression (Gil-Carvajal compactness of the auditory image) are also assessed by listeners independently of the consistent with recent data which suggest that other spatial parameters (azimuth and unaffected by congruent or incongruent visual input. On the other hand, this result is

Figure 3 reproduces the raw data used in these seven scatter-plots. A systematic effect of a specific type of audiovisual incongruence would be evident in these plots by a shift of the data points away from the identity line, with a shift towards the horizontal/vertical axis suggesting a higher/lower perceived degree of reverberation in congruent conditions, respectively. We found no type of audiovisual incongruence that was rated significantly differently to congruent conditions by the subjects.

4. Discussion

The current data show that the subjects could reliably judge the degree of reverberation of a presented virtual environment. They show equally clearly that the subjects’ reverberation judgements did not depend on whether a visual representation of an auditory environment was provided to them during listening, and that they were also not systematically affected by the congruence or incongruence of the presented auditory and visual environments.

Considering the dominance of the visual system in spatial awareness and its established influence on the perception of single echoes, it might on the one hand appear surprising that judgements of a room-related attribute of sound are entirely unaffected by congruent or incongruent visual input. On the other hand, this result is consistent with recent data which suggest that other spatial parameters (azimuth and compactness of the auditory image) are also assessed by listeners independently of the simultaneous visual impression (Gil-Carvajal et al., 2016). These authors only detected an effect of visual stimulation on sound source distance, where multi-modal integration probably assigns a larger weight to visual information due to the relatively lower reliability of auditory cues for distance perception (Loomis et al., 1998).

In this context, it might be important to note that it was easily possible for our subjects to determine the azimuthal position of the sound source not only as mediated by the visual stimulation (the position of the TV set in the virtual room), but also as mediated by the auditory stimulation. While the audio-visual stimulation technique employed in this experiment is theoretically suitable to elicit a visual capture effect in conditions where the visual source position deviates from the sound source position, this is unlikely to have taken place considering the magnitude of the azimuthal
incongruence (30° or more), which exceeds typical thresholds of perceptual fusion (Hendrickx et al., 2015).

It should not be overlooked that despite a lack of statistical significance at \( \alpha = 0.05 \), test statistics and potential effect sizes are largest when comparing conditions that modulate sound or visual source azimuth. Relatively high score differences and comparatively low \( p \)-values are observed between stimulus conditions that have a 0° vs either a 60° or a 90° discrepancy between sound and visual source azimuth, with lower perceived degrees of reverberation in the azimuthally incongruent conditions. We do not believe that this is due to audiovisual interactions, given that a better-ear effect in listening in rooms offers a simpler explanation: Because subjects always looked towards the visual source position in audiovisual conditions, one of their ears was turned towards the sound source when its location diverged in azimuth. This behaviour was found to improve spatial release from masking in a single-speaker, single-masker condition (Grange and Culling, 2016) and may well affect the perception of reverberation.

It should be noted that the initial audio only training for the judgement of reverberation might have primed a listener’s focus on the auditory cues while tending to more disregard visual cues. It is unclear whether results might be different for more naive listeners which do not receive audio only training ahead of the task and which judge perceived reverberation of the overall scenario or its effects on perception in a more indirect task.

Acknowledgments
This research was supported by the Munich Center for Neurosciences, the Bernstein Center for Computational Neurosciences Munich, the Graduate School of Systemic Neurosciences, the Deutsche Forschungsgemeinschaft (DFG) FOR 1732 (TPE) to author S.E., and the DFG Grant No. wi1518/17 to author L.W. We thank Baccara Hizli for her assistance in data acquisition. The visual stimuli depicted in Fig. 1 are based on 3D models provided by Rui Teixeira (bedroom; public domain), DragonautX on blendswap.com (office; CC-BY 3.0), and Fatih Eke (factory; CC-BY 3.0).

References and links
Bishop, C. W., London, S., and Miller, L. M. (2011). “Visual influences on echo suppression,”Curr. Biol. 21(3), 221–225.
Bishop, C. W., London, S., and Miller, L. M. (2012). “Neural time course of visually enhanced echo suppression,” J. Neurophys. 108(7), 1869–1883.
Bishop, C. W., Yadav, D., London, S., and Miller, L. M. (2014). “The effects of preceding lead-alone and lag-alone click trains on the buildup of echo suppression,” J. Acoust. Soc. Am. 136(2), 803–817.
Blauert, J. (1997). Spatial Hearing: The Psychophysics of Human Sound Localization (MIT Press, Cambridge, MA).
Blender Foundation and community (2018). Blender 2.79b—free and open 3D creation suite, https://www.blender.org/.
Brandewie, E., and Zahorik, P. (2010). “Prior listening in rooms improves speech intelligibility,” J. Acoust. Soc. Am. 128(1), 291–299.
Brown, A. D., Stecker, G. C., and Tollin, D. J. (2015). “The precedence effect in sound localization,” J. Assoc. Res. Otolaryng. 16(1), 1–28.

Clifton, R. K. (1987). “Breakdown of echo suppression in the precedence effect,” J. Acoust. Soc. Am. 82(5), 1834–1835.

Clifton, R. K., and Freyman, R. L. (1989). “Effect of click rate and delay on breakdown of the precedence effect,” Percept. Psychophys. 46(2), 139–145.

Clifton, R. K., Freyman, R. L., and Meo, J. (2002). “What the precedence effect tells us about room acoustics,” Percept. Psychophys. 64(2), 180–188.

Gil-Carvajal, J. C., Cubick, J., Santurette, S., and Dau, T. (2016). “Spatial hearing with incongruent visual or auditory room cues,” Sci. Rep. 6, 37342.

Grange, J. A., and Culling, J. F. (2016). “The benefit of head orientation to speech intelligibility in noise,” J. Acoust. Soc. Am. 139(2), 703–712.

Hartung, K., and Trahiotis, C. (2001). “Peripheral auditory processing and investigations of the ‘prece-dence effect’ which utilize successive transient stimuli,” J. Acoust. Soc. Am. 110(3), 1505–1513.

Hendrickx, E., Paquier, M., Koehl, V., and Palacino, J. (2015). “Ventriloquism effect with sound stimuli varying in both azimuth and elevation,” J. Acoust. Soc. Am. 138(6), 3686–3697.

Lindau, A., Erbes, V., Lepa, S., Maempel, H.-J., Brinkman, F., and Weinzierl, S. (2014). “A spatial audio quality inventory (saqi),” Acta Acust. Acust. 100(5), 984–994.

Litovsky, R. Y., Colburn, H. S., Yost, W. A., and Guzman, S. J. (1999). “The precedence effect,” J. Acoust. Soc. Am. 106(4), 1633–1654.

Loomis, J. M., Klatzky, R. L., Philbeck, J. W., and Golledge, R. G. (1998). “Assessing auditory distance perception using perceptually directed action,” Percept. Psychophys. 60(6), 966–980.

McCreery, A., and Calamia, P. (2006). “Cross-modal perception of room acoustics,” J. Acoust. Soc. Am. 120(5), 3150–3150.

Postma, B. N., and Katz, B. F. (2017). “The influence of visual distance on the room-acoustic experience of auralizations,” J. Acoust. Soc. Am. 142(5), 3035–3046.

Pulkki, V. (1997). “Virtual sound source positioning using vector base amplitude panning,” J. Audio Eng. Soc. 45(6), 456–466.

Satterthwaite, F. E. (1946). “An approximate distribution of estimates of variance components,” Biometr. Bull. 2(6), 110–114.

Stein, B. E., and Meredith, M. A. (1993). The Merging of the Senses (MIT Press, Cambridge, MA).

Tolnai, S., Litovsky, R. Y., and King, A. J. (2014). “The precedence effect and its buildup and breakdown in ferrets and humans,” J. Acoust. Soc. Am. 135(3), 1406–1418.

Traer, J., and McDermott, J. H. (2016). “Statistics of natural reverberation enable perceptual separation of sound and space,” Proc. Natl. Acad. Sci. 113(48), E7856–E7865.

Watkins, A. J. (2005). “Perceptual compensation for effects of reverberation in speech identification,” J. Acoust. Soc. Am. 118(1), 249–262.

Wendt, T., van de Par, S., and Ewert, S. D. (2016). “Perceptually plausible acoustics simulation of single and coupled rooms,” J. Acoust. Soc. Am. 140(4), 3178–3178.