Subjective scaling of spatial room acoustic parameters influenced by visual environmental cues

Daniel L. Valente^{a)} and Jonas Braasch

Communication Acoustics and Aural Architecture Research Laboratory (CA³RL), School of Architecture, Rensselaer Polytechnic Institute, 110 8th Street, Troy, New York 12180

(Received 21 November 2008; revised 12 July 2010; accepted 20 July 2010)

Although there have been numerous studies investigating subjective spatial impression in rooms, only a few of those studies have addressed the influence of visual cues on the judgment of auditory measures. In the psychophysical study presented here, video footage of five solo music/speech performers was shown for four different listening positions within a general-purpose space. The videos were presented in addition to the acoustic signals, which were auralized using binaural room impulse responses (BRIR) that were recorded in the same general-purpose space. The participants were asked to adjust the direct-to-reverberant energy ratio (D/R ratio) of the BRIR according to their expectation considering the visual cues. They were also directed to rate the apparent source width (ASW) and listener envelopment (LEV) for each condition. Visual cues generated by changing the sound-source position in the multi-purpose space, as well as the makeup of the sound stimuli affected the judgment of spatial impression. Participants also scaled the direct-to-reverberant energy ratio with greater direct sound energy than was measured in the acoustical environment.

© 2010 Acoustical Society of America. [DOI: 10.1121/1.3478797]

PACS number(s): 43.66.Lj, 43.55.Br, 43.55.Hy [LMW]

Pages: 1952–1964

I. INTRODUCTION

It is well known that visual cues can influence or even dominate over acoustic cues during the perceptual analysis of a multimodal scene. A classic example of this auditory-visual integration is the ventriloquist effect^{1–4}—a phenomenon directly related to spatial auditory-visual mismatch. This effect occurs in situations where synchronous auditory and visual information are presented in separate physical locations. The two events become spatially fused, and the perceived location of the auditory stimulus is pulled in the direction of the visual stimulus.

Beerends and Caluwe⁵ have shown that video quality can be more influential in shaping a subject's perceived quality of an audio-visual scene than audio quality. Also, as has been demonstrated in the McGurk⁶ effect, when participants are shown the video of a person speaking one syllable such as “ba” and presented with an audio recording of a person speaking a syllable such as “ga,” an intermediate precept syllable “da” is reported. In judgments of spatial nature, visual cues can easily dominate over acoustic cues, or sway the precept of an auditory event toward a contrasting visual one.¹

In contrast, very little has been reported regarding how visual cues influence auditory events with respect to auditory spatial parameters other than source location such as spatial impression, the focus of the study presented here. Spatial impression has been explored and defined thoroughly in previous audio-only studies.^{7–16} Early investigations by Marshall,⁷ Barron,⁸ and Barron and Marshall⁹ focused on the influence of early lateral room reflections. In 1989,

Morimoto and Maekawa¹⁰ reported that spatial impression was actually a multidimensional parameter which consisted of at least two parts: apparent source width (ASW) and listener envelopment (LEV). They described ASW as the “width of a sound image fused temporally and spatially with the direct sound image” and LEV as the “degree of fullness of sound images around the listener, excluding a sound image composing ASW.” In agreement with this, Bradley and Soulodre^{11,12} were able to correlate specific room-born reflections into the creation of ASW and LEV.

Bradley and Soulodre¹¹ also reported that early laterally arriving reflections (measured within a window of 80 ms after the arrival of the direct sound) increased ASW whereas later laterally arriving reflections that occurred after 80 ms when measured from the onset of the direct sound resulted in LEV. Further research by Morimoto *et al.* in 2001¹³ has shown that room reflections arriving from behind the listener increase LEV, and while early reflections can increase LEV, later arriving reflections are more effective. A schematic image showing how ASW and LEV are two separately-occurring perceptual dimensions of spatial impression is presented in Fig. 1, adapted from a figure of Morimoto *et al.*'s study.¹³ Figure 1 shows ASW of an auditory event as a spatial dimension that occurs in front of the listener while the LEV surrounds the listener in all directions, but is perceived separately from the ASW. In addition, Bradley and Soulodre¹¹ and Morimoto *et al.*¹³ have provided experimental results that revealed which components of the acoustical input of an enclosure can be attributed to LEV and ASW and that listeners can perceive them as two separate attributes of spatial impression. Based on previous work to test the influence of varying visual-stimulus parameters on the perception of an acoustic stimulus (Ventriloquism,^{1–4} McGurk effect⁶), the question may be then raised: Can the visual characteris-

^{a)} Author to whom correspondence should be addressed. Present address: Boys Town National Research Hospital, Omaha, NE. Electronic mail: daniel.valente@boystown.org

FIG. 1. Schematic representation of the concepts of ASW and LEV [adapted with permission from Morimoto *et al.* (Ref. 13)].

tics of an acoustical environment or the sound-emitting source modulate the judged spatial impression of that environment?

There are far fewer studies investigating cross-modal interactions in acoustical spaces than the number of previously mentioned audio-only studies on spatial impression. Larsson *et al.*¹⁷ examined perceptual room acoustical parameters in the context of vision. The investigators presented acoustical stimuli with and without a coincident visual stimulus to participants. They found that the participants reported the “distance-to-source” parameter as greater when presented with auditory-only stimuli and shorter when subjects were presented with cues in both the auditory and visual domains. Cabrera *et al.*¹⁸ gathered subjective responses from participants when they were presented audio-only and visual-only stimuli taken from the same concert hall environments. They found that the participants rated similar perceptual terms such as “envelopment” and “intimacy” differently for the audio-only and visual-only conditions.

Jeon *et al.*,¹⁹ in studying the effect of visual and auditory cues on seat preference in an opera theater, found that both auditory and visual cues were significant in preference judgments. Overall sound pressure level (SPL) augmented by visual preference influenced their participants’ judgments. The investigators used photographs projected on a screen at different seat locations in the theater and presented participants with binaural recordings of the acoustical environment at that seat. The sound-emitting source was not shown on stage, but rather a singer recorded in an anechoic chamber was convolved with a pre-recorded binaural room impulse response (BRIR).

In experiments conducted in actual auditoria, Barron²⁰ found that listeners tend to compress loudness judgments compared to what is objectively measured based on the source-to-receiver distance. In their case, participants’ expectations regarding sound source level was judged as greater than the actually measured values at larger source-receiver distances.

Inclusion of a visual representation of the performance also appears to be important when judging auditory spaciousness. Obviously, it is difficult to measure the subjective expectation with regard to a performer in a space without representing the performer visually in the environment. While the existence of audio-visual interactions was demonstrated in the aforementioned multi-modal studies, the extent to which the subjective expectation with regard to a room would change if the performing musicians were included in the visual environment remains uncertain.

Valente and Braasch²¹ and McCreery and Calamia²² addressed the issue of including a sound-emitting source in a judged audio-visual scene. These studies did not focus on spatial impression but on other room-acoustical parameters (direct-to-reverberant energy ratio, reverberation time). While reverberation time is a room-specific constant that usually does not vary significantly with location, the direct-to-reverberant energy ratio is a function of source-to-receiver distance that also depends on the source’s characteristic directivity pattern. In agreement with this, the participants in McCreery and Calamia’s experiment had a clear expectation induced by the visual cues about how the direct-to-reverberant energy ratio should vary with source position.²² Valente and Braasch²¹ also showed that the participants had a clear expectation about the reverberation time of a performance/concert venue, once they saw the location. In addition, a significant difference for expectation judgments was found between a concert harp and drums.²¹ It was hypothesized that the chosen settings for the drum contained less reverberant energy than the settings for the harp, because in commercial multi-track recordings drums are often processed with less reverberation than is the case for other instruments.

The impetus for the current study was to investigate how visual cues may impact spatial-impression expectations of an environment. As such, this study addressed the following questions: (1) Can judgments of spatial impression be modulated by the visual makeup of the sound-emitting source? (2) Will the type of visual feedback change how participants build up their expectations of how a room will sound? (3) Will participants change their expectations of direct-to-reverberant energy ratios and judged ASW and LEV if an empty room is shown instead of a performer playing in that room? Further, it was of interest to examine the extent to which spatial impression varies for different types of musical instruments. In the experiment, a snare drum, a djembe, a violin, a theremin, and spoken voice were used to test if similar observations would be made with respect to spatial impression. It was also of interest to examine how the speech stimulus was rated with respect to the other percussive and non-percussive stimuli.

II. METHOD

A. Multi-purpose room

Acoustical measurements and video recordings were conducted in a large, multi-purpose space, the Greene Gallery, located at the School of Architecture at Rensselaer Polytechnic Institute. This space was chosen, in part, for its even

FIG. 2. (Color online) Source and receiver positions within the multi-purpose space that was used to record the BRIRs and background video recordings (plan). The source, labeled S1, was placed in the middle of the 7.5-m by 25-m space, 2-m into the room from the back wall. The binaural dummy head was placed in a 6-m $\times$ 2-m grid, 6 m from the position of the source. Four receiver positions were used for the experiment, labeled R1, R2, R3, R4.

lighting, low background noise (L_{eq} 15 min $<$ 30.0 dB SPL A), uniformity in construction from front to back, and a reverberation time close to that of a typical performance space intended for Western classical music (T_{30} Mid: 1.8 s). The space was a rectangular room 7.5 m $\times$ 25 m $\times$ 8 m, length $\times$ width $\times$ height. The space was primarily of wooden construction with non-full-height (4 m) moveable Homasote panels in the rear of the space and fixed, full-height Homasote paneling on the side and front walls.

B. Source and receiver positions

A plan view of the gallery space, as well as the positions of the sources and receivers used for the experiment are shown in Fig. 2. BRIRs were measured at four receiver po-

sitions labeled: R1, R2, R3 and R4. Position S1 corresponds to a single source position on the midline of the long dimension of the space 2 m from the rear wall. Position R1 corresponds to a listening position at a distance of 6 m from the source. Position R2 corresponds to a position 6 m from the source and 2 m to the right of the source. Position R3 corresponds to a position that is 12 m from the source. Position R4 corresponds to a position 12 m from the source and 2 m to the right of the source.

BRIRs were measured using a diffuse-field equalized binaural dummy head (Head Acoustics, HMS III.L) positioned at one of the four measurement locations. The dummy head was positioned on a chair next to an HDV camera (Canon, HV20), which captured coincident visual stimuli at the given receiver position.

The responses were measured using a 50-s logarithmic sinusoid-sweep excitation signal with a 24-bit resolution analog-digital (A/D) conversion and 48-kHz sampling rate.

In order to generate the measured impulse, the recorded sweep was convolved in the time domain using a proper time-inverse filter with -6 dB/Octave equalization to compensate for the logarithmic frequency of the excitation signal. This method is described by Farina²³ and allows the measurement of impulses with a very large signal-to-noise ratio (SNR) as required for auralization.

The SNRs of the measured BRIRs were above 70 dB. The measurement loudspeaker used was a three-way omnidirectional unit consisting of a commercially available subwoofer (Eastern Acoustic Works, SB120R), and a mid-range aluminum-cast dodecahedron loudspeaker with full range, 8-cm diameter transducers (Audax, AP080G0 speakers). In order to extend the high-frequency cutoff of the omnidirectional radiation pattern of the measurement loudspeaker, a custom-designed dodecahedron tweeter was employed. The addition of the tweeter-array yields a standard deviation of the polar radiation that is less than 5 dB from 4 kHz to 8 kHz and less than 6 dB from 8 kHz to 20 kHz. The subwoofer unit was driven from 50 Hz to 250 Hz, the mid-range dodecahedron loudspeaker was driven from 250 Hz to 5 kHz, and the high-frequency dodecahedron loudspeaker was driven from 5 kHz to 20 kHz.

C. Measurement results

The measured reverberation time, analyzed in octave bands for the four listening positions, can be seen in Fig. 3. The reverberation time was quite uniform at each receiver location, varying by a maximum of 0.07 s at 1 kHz between R1 and R2. The T_{30mid} was 1.82 s for R1, 1.86 s for R2, 1.83 s of R3, and 1.86 s for R4. C_{7mid} values were -14.3 dB for R1, -13.0 dB for R2, -16.6 dB for R3, and -19.2 dB for R4. C_{7mid} was chosen to express the relative change in the direct-to-reverberant energy ratio (D/R ratio) across the measurement positions. C_7 is defined as:

$$C_7 = 10 \log_{10} \left(\frac{E_7}{E_\infty - E_7} \right) \text{dB}, \quad (1)$$

whereas E_7 denotes sum of the energy in the BRIR from the onset of the direct sound to 7 ms. The ratio is determined by

FIG. 3. The measured reverberation time analyzed in octave-bands for the four listening/viewing positions in the performance space used for this experiment. The reverberation time (T_{30}) at the measurement positions (labeled R1, R2, R3 and R4, see Fig. 2) has been calculated in octave bands, based on the recommendations of ITU-1997 (Ref. 25).

dividing the energy by the difference between the later energy (after 7 ms) and the E_7 . For C_{7mid} , the average of the C_7 in the octave bands of 500, 1000 and 2000 Hz is taken.

In addition, the binaural quality index (BQI) was determined for each measurement position. For this study, the BQI definition of Hidaka *et al.*²⁴ was followed. Here the normalized inter-aural cross correlation magnitude (IACC), which is the interaural coherence, is calculated from the early component of the BRIRs (0–80 ms as measured from the onset of the direct sound). The calculation is performed for different octave bands with center frequencies of 500, 1000, and 2000 Hz. The BQI is then defined as the negative IACC plus one (1-IACC). The BQI was higher in positions further away from the sound source where R3 was 0.81 and R4 was 0.80. The receiver positions closer to the sound source had lower BQIs (R1, 0.62 and R2, 0.77).

Finally, the 1-IACC late objective parameter derived from the measured BRIR (from 80 ms after the direct sound to the end of the BRIR) was calculated for each receiver position. The 1-IACC values were quite uniform; at R1, R3, R4, the 1-IACC was 0.86. At R2, the 1-IACC was 0.84.

D. Sound stimuli

The sound stimuli captured for the experiment consisted of a snare drum, a djembe, a violin, a theremin, and a spoken voice. Performances were recorded in a dark-walled, sound-treated studio (Rensselaer Polytechnic Institute's NYSTAR Experimental Telepresence and Virtual Acoustics Laboratory, height $\times$ width $\times$ length, $3 \times 4 \times 5$ m³). The room walls were treated with thick insulation, and absorptive panels were hung from the ceiling to reduce echoes and reflections. The measured acoustic parameters of the recording studio are found in Table I. By using close-microphone techniques and recording the performances in a room with a very short T_{30mid} value (160 ms at 1 kHz), the aim was to maximize the D/R ratio so that the recording room would not confound the experiment by adding additional reflected energy that would not have been present in the multi-purpose space. The C_{7mid} value for the recording studio was 2.2 dB (measured using

TABLE I. Octave-band acoustic parameters of the studio used to record the stimuli for the experiment.

Parameter	Octave-band center frequency					
	125 Hz	250 Hz	500 Hz	1000 Hz	2000 Hz	4000 Hz
C50 (dB)	19.54	22.08	28.96	22.61	26.80	25.19
C80 (dB)	28.81	32.04	38.91	30.36	34.77	34.34
EDT (s)	0.14	0.13	0.10	0.04	0.06	0.19
T30 (s)	0.18	0.17	0.12	0.16	0.15	0.11

the same techniques as described for the acoustical measurements taken in the multi-purpose space). This measurement was taken with the excitation sound source located where the sound stimuli were to be recorded and the measurement microphone positioned 2 m from that source.

The visual components of the performances were recorded at a distance of 3 m on a blue-screen background (Reflecmedia, Chromatte Small Studio) using an HDV camera (Canon, HV20). The background covered both the area behind and beneath the musician.

An omni-directional condenser microphone (Earthworks, QTC30) was used to record the acoustic performance using close-microphone techniques to minimize any reflections in the reference room. A low-noise microphone pre-amplifier and digital-audio converter was used for the audio-stimulus acquisition (Apogee, Duet). The stimuli were recorded into a digital-audio workstation (DAW) software package (Apple, Logic Pro 8), and A-weighted Sound Pressure Level (SPL) measurements were taken at a distance of 1 m from the source using a type-2 SPL meter (Ono Sokki, LA-4240). At the end of the recording session, the five musicians were presented with both the auditory and visual performances. All of them verified that their performances sounded and looked natural to them.

E. Digital compositing for visual source positioning

In order to reduce bias and variability between samples in the experiment, the method of recording a performer *in-situ* for each to-be-tested audio-video scene was rejected in favor of a digital-compositing (visual) and real-time convolution (acoustic) method.

For the visual component of this method, each performance was recorded once in the studio and then positioned, scaled and shaded into the multi-purpose space's environmental video recording using digital matting. This technique allows the performer to appear in the actual location of the music-performance environment.

Digital matting is the process of extracting a foreground object from a background with the purpose of superimposing it onto a secondary background. This latter process is known as compositing and was used in this experiment to create a new setting for the source performances. The technique used for this study has been described by Smith and Blinn²⁶ and is called *Blue-Screen Matting*. The performer is captured against a monochromatic background (either blue or green). By *keying* or sampling the color of the background and by setting the *alpha* or opacity to a value between zero and one, a clean matte of the foreground image can be pulled out (the

FIG. 4. The compositing technique is shown from single-frame images within the multi-purpose space. Frame A shows the image captured within the space. Frame B shows the result of the compositing process, where a properly scaled video recording of a musician is superimposed in the room.

pure blue or green color will not be found on the foreground image, and thus the latter will not be made transparent). After a suitable alpha mask is created of the foreground, it can be placed on top of a secondary background, utilizing the operation known as the *compositing equation*:

$$C = \alpha F + (1 - \alpha)B, \quad (2)$$

where C represents the pixel composite, F the foreground, B the background and α the opacity, which add up to one at each pixel. Examples of the matting can be seen in Fig. 4, where the musical performance video recording in the recording studio is positioned within a video recording taken in the multi-purpose space at receiver location: R1.

F. Participants

Twelve participants, ages 23–50 (median age 34), were tested in the experiment. Five participants were female and seven were male. Seven of the participants had experience in acoustic design or signal processing, and five of the participants had formal training in a musical instrument. Four of the participants had experience in both acoustics and musical performance. Participants were recruited on a volunteer basis and did not receive compensation. Participants were naïve to the hypothesis under test and self-reported that they had no hearing or visual disorder that would prevent them from making the judgments that they would be asked to perform during the test. Informed consent procedures were approved by the Institutional Review Board at Rensselaer Polytechnic Institute.

G. Independent variables

Three independent variables were chosen for the experiment: sound stimuli (5-level), visual stimuli (4-level) and receiver position (4-level). The sound stimuli consisted of four performances of music and one of speech. These were previously recorded and edited to create 30-s clips. The music performances involved a snare drum, a djembe, a violin, and theremin (labeled A1, A2, A3, and A4) while the speech involved sentences chosen from the IEEE/Harvard Speech

FIG. 5. Images of the sound stimuli used. The music performances involved a snare drum, a djembe, a violin, and theremin (labeled A1, A2, A3, and A4) and the speech stimulus is labeled A5.

Intelligibility Corpus²⁷ (labeled A5). Images of the sound stimuli composited in the multi-purpose room can be seen in Fig. 5.

The visual stimuli represent the appearance of the visual cues for each audio-visual scene. Four visual cues were provided: none, surrogate, coherent, and additive, labeled V1, V2, V3 and V4. In the case of *no cue*, the condition contained an empty listening room and the acoustic performance only. The *surrogate* condition consisted of the image of a 0.25-m tall loudspeaker positioned on a 1.25-m tall loudspeaker stand in the listening room. The *coherent* visual condition consisted of the video recording of the musical or speech performance positioned in the listening room. The *additive* condition consisted of the video recording of the musical or speech performance, with the addition of two 0.25-m tall loudspeakers positioned 1 m to the right and left of the performing musician, on 1.25-m tall loudspeaker stands. The visual stimuli predictor is ordered in terms of increasing visual width with the hypothesis that participants will rank ASW greater as a function of increasing visual width, and perhaps augment their scaling of LEV due to their expectations regarding the source's angular directionality. The receiver position represents the judged audio-visual scene presented in Fig. 2. The location of the receiver positions were positions R1: near-center, R2: near-right, R3: far-center, and R4: far-right.

H. Dependent variables

The BRIRs measured in the same environment (Greene Gallery) were used to reproduce the multi-purpose space as accurately as possible. In the experiment, the participants were asked to adjust the levels of the direct sound source and reverberant tails of the presented performance to meet their expectations after being exposed to the visual environment. The adjustment of other important parameters such as reverberation time and initial time delay gap, would not have been feasible with measured BRIRs. The acoustical stimuli were created by convolving a single-channel audio source with measured BRIRs that corresponded to each receiver location. The BRIR was split temporally at the point of direct-sound

FIG. 6. Schematic diagram showing how the measured BRIR at a receiver location was split temporally after the offset of the direct sound (S_d). By convolving the measured direct sound and reverberant tail (S_{rev}) independently with a sound stimulus, participants could mix the levels to artificially increase (c), decrease (a) or maintain the energy ratio of the direct sound (E_{direct}) to exponential decay (E_{reverb}), (b). The reverberation time (T_s) was based on the receiver position of the measurement.

offset (based on the physical distance between the source and the receiver) so that participants could mix the levels of direct and reverberant energy independently during the test as seen in Fig. 6.

Four different types of responses were collected from the participants: two rankings of spatial impression and the early and late energy levels of the impulse response. First, there was the subjective ranking of apparent source width, followed by the subjective ranking of listener envelopment. Both responses corresponded to a continuous range on a slider, the value of which was not reported to the participant during the experiment. For analysis, the slider values were encoded using a linear scaling function from 0.1 to 7.0 in steps of 0.1 units.

In addition to the ranking of the spatial impression of each judged scene, participants were asked to match the direct and reverberant energy levels of the split BRIR to match their expectations of how the environment would sound at a given receiver position. The first component collected was the direct sound-energy level. This was the amount of sound energy transmitted along the direct path from source to receiver. The experiment was configured so that the maximum possible signal level was equal to a peak level of 94 dBA (for the snare-drum stimulus), which was calibrated to 0 dB full-scale (dBfs) at a distance of 1-m from the source. The second component collected was the reverberant sound-energy level, the amount of sound energy present in the summation of the reflected path from source to receiver. It was configured in the same way as the direct sound component. For every audio-visual scene, participants were able to adjust the direct sound-energy level and the reverberant sound-energy level of the presented performance from -90 dBfs to 0 dBfs. The numeric value of the level was not shown to the participant who was able to adjust the response continuously using a slider.

For analysis, the differences between the direct and reverberant components reported were gathered to determine how participants expected the multi-purpose space to sound relative to the measured BRIR. Placing the direct sound component greater than the reverberant component would result in a positive D/R ratio mismatch (x more dB would be

expected by the participants vs. what was found in the original recorded BRIR). Placing the direct sound component less than the reverberant component would result in a negative D/R ratio mismatch. This ratio was used for further analysis particularly given previous results by McCreery and Calamia,²² and by Valente and Braasch,²¹ who found that participants routinely expected a greater D/R ratio compared to what was found in the actual acoustical environment.

I. Procedure

Before the test began, an orientation session was undertaken with each participant. This session consisted of brief equipment training, a pre-interview about their musical and acoustics background and a definition of the response parameters being collected. For this test, ASW and LEV were explained to the participant as follows, based on definitions by Bradley and Soulodre:^{11,12}

“Apparent source width (ASW) is the apparent auditory width of the sound field created by a performing entity as perceived by a listener in the audience area of a concert hall.”

and

“Listener envelopment (LEV) is the subjective impression by a listener that (s)he is enveloped by the sound field, a condition that is primarily related to the nature and density of late-arriving lateral reflections.”

Participants were seated in a chair positioned in the center of the same studio that was used to record the sound stimuli, and they were asked to listen to the binaural stimuli with electrostatic headphones (open-air Stax Lambda Pro with diffuse-field equalization). No head-tracking device was used.

For the experiment, the visual stimuli were reproduced on a projection screen using a projector capable of life-size reproduction of the recorded stimuli at the reference listening position. The native resolution of the projector was 1280×1024 pixels, which ensured a quality image with minimal visual distortion, and supported the high-resolution, 720p video standard (1280×720 pixels). The resulting auditory and visual stimuli were processed in real-time using a custom-designed Max/MSP/Jitter²⁸ patch where the desired background image and lighting conditions could be synthesized for each individual stimulus.

A flow chart describing the experimental process can be seen in Fig. 7. For each audiovisual scene, the custom-designed software patch (Max/MSP/Jitter) loaded the movie file (Quicktime format). The movie file contained the visual stimulus and the multi-channel audio signal. The audio signals were stored separately for the direct signal and the signals containing the room reflections. The separation was achieved by splitting the BRIRs into the direct-sound component and the remaining part, and then convolving both components separately with the closely recorded sound of the musical instruments/voice. For the procedure—which allowed the mixing of the audio signal in real-time according to the level adjustments of the participants—the BRIRs were taken from the same location and angle that was used to

FIG. 7. (Color online) Experimental flow chart. This describes the procedure used by the participant for scaling each test audio-visual scene.

record the visual stimulus. The audio signals were played back from a computer using a sound card at 48-kHz sampling frequency (Apogee Duet Firewire interface). The stimuli were assessed by the participants through the use of a MIDI control surface (Behringer, BCF2000). The control-surface faders were continually variable and had no default “release values;” when a given fader was released, the position was held until further adjustment was made by the participant. Four 100-mm faders were assigned to adjust the values of the four response types (direct energy level, reverberant energy level, ASW, and LEV). Moving the first fader resulted in a change in the level of direct sound-energy of the auditory presentation and moving the second fader resulted in a change of the level of reverberant sound. Moving the third and fourth faders did not result in any change of the audio rendering, but were used to capture the perceived levels of ASW and LEV, respectively.

In each experimental condition, the participant was presented with an audio-visual performance, positioned the levels of the four faders, and finally stored the values, which automatically advanced the experiment to the next condition. Between each tested scene, a five-second pause was used, where the acoustic and visual stimuli were zeroed-out in order to mitigate weighting effects between two adjacent test scenes. Each complete test took between 75 and 90 min. Participants were encouraged to take as many breaks as they wished. The participants were allowed to listen to the 30-s performance clip as many times as they wished before storing the values for any given audio-visual scene. Before the start of each experimental set, the participant was instructed to explore the limits of the control surface on the acoustic rendering model, but was not explicitly instructed as to the range of the controllable parameters.

Each participant’s test consisted of three complete repetitions of the treatment combinations described in Table II.

TABLE II. Matrix of treatment combinations for the experiment. Cond. number is the index of the condition being addressed by the participant. Auditory stimuli denotes the musical or speech performance that is played; A1: Snare Drum, A2: Dejembe, A3: Violin, A4: Theremin, A5: Speech. Visual stimuli denotes what the participant sees visually in the testing room; V1: no visual stimulus, V2: loudspeaker (surrogate), V3: coherent (actual video of the performer), V4: additive (video of the performer and 2 loudspeakers positioned to the left and right of the performer). Location (Loc.) is the index of the physical receiver location in the judgment room where the participant is making scaling judgments.

Cond. number	Auditory stimuli	Visual stimuli	Loc.
1	A3	V1	R2
2	A2	V1	R1
3	A2	V1	R4
4	A4	V1	R1
5	A5	V1	R2
6	A1	V1	R1
7	A3	V1	R4
8	A2	V1	R3
9	A5	V1	R4
10	A4	V1	R3
11	A5	V2	R3
12	A4	V2	R4
13	A3	V2	R2
14	A5	V2	R2
15	A2	V2	R1
16	A3	V2	R4
17	A5	V2	R1
18	A4	V2	R3
19	A1	V2	R1
20	A1	V2	R4
21	A2	V3	R2
22	A3	V3	R3
23	A3	V3	R2
24	A5	V3	R3
25	A5	V3	R2
26	A1	V3	R3
27	A4	V3	R2
28	A2	V3	R2
29	A4	V3	R4
30	A1	V3	R4
31	A3	V3	R1
32	A2	V4	R4
33	A1	V4	R1
34	A5	V4	R1
35	A5	V4	R4
36	A1	V4	R2
37	A2	V4	R3
38	A3	V4	R3
39	A3	V4	R1
40	A1	V4	R3
41	A4	V4	R1
42	A2	V4	R2
43	A3	V4	R4
44	A4	V4	R2

For each block of trials the treatment combinations were randomized. These treatment combinations created composite audio-visual scenes that participants scaled.

J. Data analyses

Data were analyzed to address the research questions posed in the introduction. A linear mixed-model was fit sepa-

rately to each of the three dependent variables (ASW, LEV and D/R ratio) using statistical design software (JMP 8.1, SAS Institute²⁹), with subject and repetition as random effects and three fixed effects, visual stimuli (4 levels), sound stimuli (5 levels) and receiver position (4 levels). Random effects were calculated using the Restricted Maximum Likelihood (REML) method in JMP,³⁰ which uses the Kenward-Roger method³¹ to calculate “effective” denominator DFs that correct p-values such that they do not over-reject the null hypothesis for each fixed effect.

Paired-samples Tukey-Kramer HSD or “Honestly Significant Difference” tests were also performed *post-hoc* for each significant predictor’s least-squares mean level. Significance levels were set *a priori* at a level of 0.05.

An analysis of the repetition on participants’ responses for each of the dependent variables revealed that repetition accounted for less than 1.5% of the variance for each of the fit linear mixed-models, so for further analysis the mean response of each participants’ three repeated responses for a given trial was taken.

III. RESULTS AND DISCUSSION

A. Apparent source width

It was hypothesized that the visual characteristics of the sound-emitting source would modulate participants’ expectations of ASW. The predictor variable: visual stimuli were arranged such that the categorical levels had different widths. In the case of no visual stimuli there is no inherent visual cue regarding the dimensions of the sound-emitting source in the room. The single loudspeaker was presented as being 0.25 m wide (being narrower than the sound stimuli video recording). The loudspeaker visual stimuli may also change a participants’ expectations regarding the source directivity pattern and therefore perhaps the ASW. Finally the visual stimuli where the video recording was presented with loudspeakers to the left and right of the sound stimuli may influence the participants’ expectations regarding the ASW.

In addition, the location of the participant in the room as well as the sound stimuli should affect the ASW. A prior study by Becker and Sapp³² demonstrated that the nature of the sound stimuli could affect a participant’s judged ASW, namely music sound stimuli were reported with a greater ASW than speech or noise sound stimuli. (Panel b of Fig. 4 in Ref. 32). The current study further broke sound stimuli down into two percussive instruments, two sustained instruments and a speech sample for the presented sound stimuli.

The linear mixed-model fit yielded significant fixed effects for all three predictor factors. The visual stimuli [$F_{(3,497)}=14.55$, $p=0.0001$], sound stimuli [$F_{(4,497)}=58.49$, $p<0.001$], and receiver position [$F_{(3,497)}=7.28$, $p<0.001$] were all found to be significant in modulating participants’ ranking of ASW in the tested audio-visual scenes. A variance component estimate indicated that 55.6% of the variance comes from subject variability while 44.4% was determined by the fixed effect predictors. To determine if significant differences between categorical values support the hypothesis regarding the tested fixed effect predictors, the effects leverage procedure was used (Fit Model: effects

FIG. 8. Results for ASW for each of the three tested predictors. (a) Receiver position; (b) Visual stimuli; (c) Sound stimuli. Error bars denote standard error of the means.

leverage, JMP, SAS Institute²⁹) which demonstrates how a predictor is representing the collected data and if it is statistically significant with respect to the group mean. As all three predictors were previously shown to be significant in influencing ASW, it was important to determine which categorical levels within each predictor effect are significantly different from one another by means of a Tukey-Kramer HSD test.

Figure 8 shows the results of the least square means from the linear mixed-model fit of the collected ASW. The error bars denote the standard error of the mean. In Fig. 8(a), the influence of receiver position can be seen over and above the other two independent predictors. The ASW was rated wider from receiver position R2 to R3 ($p<0.001$) and reductions in scaled ASW from R1 to R2 as well as R3 to R4 were significant as well ($p<0.01$). As the projected distance was different at each location as well as the angle of view, it is not surprising that a participant’s expectations of ASW are influenced by receiver location. Comparing this to results derived from measurements in the multi-purpose space, a positive correlation with the binaural quality index (BQI) values was found. This positive correlation between the measured ASW and BQI values reaches significance ($R^2=0.16$, $p<0.01$) and indicates that perhaps cues inherent in the individual BRIRs early-arriving lateral reflections may have influenced the ASW rated in each receiver position. Indeed the positions with the highest BQI (R3, 0.81 and R4, 0.80) were rated with a higher ASW, and positions with lower BQIs (R1, 0.62 and R2, 0.77) were rated with a lower ASW.

It is not surprising that participants judged ASW differently while in different receiver locations. This is supported by Beranek,³³ who found in his survey that ASW varied with different seating positions in a hall or opera house. The judgment differences seem to also be influenced by the viewing angle. Positions in the middle of the room (R1 and R3), were judged with a greater ASW than those positioned off-axis and near a wall at the same distance to the sound source. The angle of view in the on-axis receiver locations allowed participants to see both walls whereas the field-of-view in the off-axis positions limited a participant’s view to the single adjacent wall. It seems that participants used visual cues taken from the receiver location as well as being influenced by laterally arriving reflections (as indicated by participants ranking ASW greater as a function of an increasing BQI).

Figure 8(b) shows the effect of visual stimuli on rated ASW. It was hypothesized that the ASW could be influenced by the visual width of the sound-emitting source. A significant increase of ASW was seen between V2 and V3 ($p < 0.001$; loudspeaker, sound stimuli recording) followed by a significant increase between V3 and V4 ($p < 0.001$; video recording of the sound stimuli coupled with loudspeakers to the left and right of the sound stimuli). The scenes with no visual stimuli were rated significantly higher than those with a loudspeaker (V2; $p < 0.001$) but not significantly different than V3 or V4. It would seem participants' expectations regarding ASW were modulated by their expectations of the directivity of a loudspeaker. It was also interesting that the participants' expectations regarding ASW could be modulated by making the sound-emitting source look wider (V4); even though the same BRIR was used in each case; a wider-looking sound source caused participants to increase their ranking of the acoustic ASW.

Figure 8(c) shows the effect of changing the sound stimuli. With regard to ASW, A5 (speech) was scaled with significantly less ASW than A1-A4 ($p < 0.01$). The scaled ASW based on the other four stimuli: A1-A4 were not significantly different from one-another ($p > 0.05$). As has been previously reported by Becker and Sapp,³² musical stimuli seem to broaden the rated ASW as opposed to a speech stimulus despite being convolved with the same BRIR. Becker and Sapp speculated that "test persons judge a virtual orchestra playing in front of them and not a signal coming from a relatively closely located speaker." In the current experiment, participants could see the visual makeup of the sound emitting source (it would be more difficult to imagine a much wider visual image than what was presented in the video recording) and still they rated the ASW higher for musical stimuli. Finally, a non-significant difference between scaled ASW was seen between percussive (A1, A2) and non-percussive (A3, A4) stimuli ($p > 0.05$).

B. Listener envelopment

It was hypothesized that the visual characteristics of the sound-emitting source would modulate participants' expectations of LEV. The predictor variable: visual stimuli may impact a participant's expectations of LEV at a given listening position. Morimoto¹⁰ indicates that the degree of IACC or later arriving lateral reflections (80 ms after the direct sound and greater) relate to LEV. Work by Bradley and Soulodre^{11,12} also supports this assertion by mentioning that the relative level of late-arriving lateral energy is a physical measure of LEV. It was hypothesized that visual cues inherent in the make-up of the sound-emitting source may influence a participant's ranking of LEV. In the case of V2 (single loudspeaker), the expectations regarding a focused directivity pattern may cause participants to expect less reflected energy to arrive laterally and judge LEV lower. Also in the case of V1, where there is no inherent visual cue of the sound source or V4 where the visual width of the source is much larger than the acoustic source measured in the multi-purpose space, participants may rank LEV higher due to the perceived arrival of more late lateral energy. Finally, as seen

FIG. 9. Results for LEV for each of the three tested predictors. (a) Receiver position; (b) Visual stimuli; (c) Sound stimuli. Error bars denote standard error of the mean.

in the judgment of ASW, the characteristics of the sound stimuli may impact the ranking of LEV. Percussive instruments (A1, A2; snare drum and djembe) have greater high-frequency transient energy—which participants pick out as individual spatial reflections as opposed to the sustained sources (A3, A4, A5; violin, theremin and spoken voice), which might cause LEV to be rated lower due to temporal smearing effects.

The linear mixed-model fit yields significant fixed effects for all three predictor factors. The visual stimulus [$F_{(3,497)}=40.0$, $p < 0.001$], sound stimuli [$F_{(4,497)}=10.14$, $p < 0.001$], and receiver position [$F_{(3,497)}=27.49$, $p < 0.001$] were all found to be significant in modulating participants' ranking of LEV in the tested audio-visual scenes. A variance-component estimate indicates that 46.3% of the variance comes from subject variability while 53.7% is determined by the fixed-effect predictors.

Figure 9 shows the results of the least square means from the linear mixed-model fit of the collected LEV. The error bars denote the standard error of the mean. In Fig. 9(a), the influence of receiver position can be seen. Participants rated LEV significantly higher in positions further away from the sound source (R3, R4) than in positions closer to the sound source (R1, R2). These differences in scaled LEV were significant in *post-hoc* tests ($p < 0.001$). A non-significant difference in scaled LEV between R1 and R2, as well as R3 and R4 was found ($p > 0.05$). Participants were able to see more of the multi-purpose space when looking at the room from the further positions, which could have influenced their scaling of the parameter.

Cues regarding the volume of the enclosure appear to influence a participant's scaling of LEV. This distance dependence of scaled LEV is in good agreement with studies on auditory distance perception. Brown and Yost,³⁴ for example, have asserted through a psychophysical experiment, that visual cues of a sound-emitting object (a noise-burst through a loudspeaker in their case) can influence the relative judgments of the auditory depth of the object by the participant. In their experiment, the authors found that auditory-depth perception was inferior in a darkened room compared to a room in which the participant was able to see the location of the source of the acoustic stimuli. This finding is consistent with the fact that auditory-distance perception is quite poor, and visual cues, when available, are used as the primary

determinants of the depth of an assessed stimulus in a given environment (Nathanail *et al.*³⁵). The relationship between the findings in the present study and those of Brown and Yost³⁴ is that visual cues inherent in seeing more of the volume of the room caused participants to rank LEV higher despite a similar reverberation time in all four receiver positions.

In Fig. 9(b), the influence of visual stimuli can be seen. Participants used visual width cues in their ranking of LEV. A significantly higher ($p < 0.001$) LEV was reported when subjects saw a wide visual source (V4) than when imagery of a single loudspeaker (V2) was seen. A significant difference in LEV was also found between V1 and V4 ($p < 0.01$). A nonsignificant difference was found in scaled LEV between V2 and V3 ($p > 0.05$). Expectations regarding how the sound-source would surround the participant in the case of V4 caused an increase in the ranking of LEV. By seeing visual imagery of the sound source that appeared narrow (V2, V3), participants rated LEV lower despite using the same recorded BRIR. In the case of no visual cue, where the participant only heard the sound-emitting source and did not see him/her, a wider scaled LEV value was reported. This seems to indicate that using a visual representation of the coherent sound source results in a lower rated LEV for the same BRIR.

In Fig. 9(c), the influence of sound stimuli can be seen. LEV was rated significantly lower for the speech stimuli (A5) than for the other sound stimuli ($p < 0.001$). In addition, the theremin (A4) was scaled with significantly ($p = 0.01$) more LEV than the djembe sound stimulus (A2). A non-significant difference in the scaling of LEV between A1, A2 and A3 ($p > 0.05$) was found. Regarding LEV, the results indicate that not only the visual characteristics of the room seem to modulate the rating of the parameter but also the sound-stimulus itself. The visual display of an enclosure influences the listener's acoustic expectations of a sound stimulus. Participants are influenced by the visual cues of the environment, which is also supported by the finding of Cabrera *et al.*¹⁸ This current investigation also finds that the visual characteristics of the sound sources influence a participant's relative spatial impression in an environment.

A correlation was found between the 1-IACC late objective parameter derived from the measured BRIR (from 80 ms after the direct sound to the end of the BRIR) and participants' mean rated LEV ($R^2 = 0.19$, $p = 0.001$). Participants in the experiment used more than the late arriving energy difference between receiver positions to base their expectations regarding LEV. This effect manifests itself through the significant differences between type of sound stimulus as well as the visual cues of the room and sound stimuli.

Finally, when comparing the trends seen in scaled ASW and LEV it was of interest to see whether the subjects' ASW judgments could be used to predict their LEV judgments. Indeed, a moderate positive correlation between ASW and LEV was found ($R^2 = 0.42$, $p < 0.001$). Previous research by Morimoto and colleagues^{10,13} as well as Bradley and Souloudre^{11,12} has shown that ASW and LEV can be identified as two separate dimensions of spatial impression. While the subjects in the current study may have shown a higher cor-

relation between ASW and LEV than the cited studies, the ability for subjects to fully differentiate the dimensions of ASW and LEV may be influenced by a number of parameters that remained constant during the study such as the room characteristics and room size.

With regard to differences between ASW and LEV, it was found that subjects scaled ASW significantly different between on- and off-axis positions at the same distance [see Fig. 8(a)]. In addition, off-axis positions (R2, and R4) were scaled with less ASW than the on-axis positions. This significant trend was not seen in the LEV results, as a significant difference in scaled LEV was only found for distance [see Fig. 9(a)]. This indicates that while the two measures are partially correlated, subjects still scaled the measures independently from one-another in certain scenes, and perhaps had independent criteria for setting the values based on what they heard and saw within a trial.

C. Direct-to-reverberant energy ratio mismatch

In addition to collecting the judgment of spatial impression for each audio-visual scene, it was of interest to determine how the D/R ratio was adjusted by participants to match the visual characteristics of the sound source and location within the multi-purpose room. The mismatch between the participants' D/R ratio adjustment and the D/R ratio measured in the actual BRIR was determined for each trial. The latter corresponds to the C_7 ratio, which was presented for each position previously in Section II C. The mismatch was measured in decibels, and any deviation by participants from 0 dB indicated that their expectation between the acoustic and visual cues was skewed from those derived from the measured BRIRs. A positive mismatch level indicated that the expected D/R ratio was larger than the one actually measured for the given condition.

Previous results by Valente and Braasch²¹ also indicated that participants' expectations regarding the D/R ratio for a given visual imagery of the acoustic environment can result in a mismatch; participants expected more direct energy for a given environment than what was found in actual measurements in the space. Valente and Braasch²¹ compared expectations for a single receiver in multiple spaces. The current study explores the possibility of changes in D/R ratio mismatch in multiple receiver positions in the same space. It was hypothesized that participants would have a positive bias with respect to D/R ratio mismatch, but that the bias would not change with varying receiver positions. This is due to the fact that a change in energy ratio is already present in the recorded BRIRs as indicated in the reduction of C_7 values as a function of distance. Therefore any bias that participants have should remain relatively constant through judging the four receiver positions used in the experiment.

Valente and Braasch²¹ also found varying expectations for a concert harp and drum set with respect to reverberation time; participants required a longer T_{30} value in the case of the concert harp to match their expectations of the room they were judging. In the experiment reported here, the T_{30} was fixed by the BRIR used at each location and the reverberation time was quite uniform at all four measured locations. It

FIG. 10. Results for D/R ratio expressed in dB for each of the three tested predictors. The D/R ratio is expressed referenced to the measured D/R ratio found in the multi-purpose space. Positive values denote a positive D/R ratio mismatch relative to the measured BRIR. (a) Receiver position; (b) Visual stimuli; (c) Sound stimuli. Error bars denote standard error of the mean.

was hypothesized that participants may attempt to increase the D/R ratio in the two percussion samples (A1, A2) to decrease the audibility of the reverberant tail, as was reported in a previous experiment by Valente and Braasch.²¹ Finally, for varying the characteristics of the presented sound stimuli, the expected directionality of the loudspeaker and coherent sound source stimuli (V2, V3) was hypothesized to modulate a participant's expectation of D/R ratio. Not seeing the sound-emitting source (V1) or expectations regarding addition of a wide sound-emitting source (V4) may cause the participant to add in more reflected energy such that the D/R ratio mismatch becomes closer to what was measured in a given receiver location.

The linear mixed-model fit yielded significant fixed effects for all three predictor factors. The visual stimuli [$F_{(3,497)}=4.85$, $p<0.01$], sound stimuli [$F_{(4,497)}=71.22$, $p<0.001$], and receiver position [$F_{(3,497)}=5.60$, $p<0.001$] were all found to be significant in modulating participants' ranking of D/R ratio mismatch in the tested audio-visual scenes. A variance component estimate indicates that 71.8% of the variance comes from subject variability while 28.1% is determined by the fixed effect predictors.

Figure 10 shows the results of the least square means from the linear mixed-model fit of the collected D/R ratio mismatch. The error bars denote the standard error of the mean. In Fig. 10(a), the influence of receiver position can be seen. It was expected that there would be no significant effect of receiver position on the judgment of D/R ratio mismatch, but a clear significant difference is seen in the pattern of the mean responses with respect to the distance from source to receiver. Participants' scaling of D/R ratio is significantly higher for the close receiver positions (R1 and R2) than for the far receiver positions (R3, R4). *Post-hoc* testing reveals that the D/R ratio mismatch is significantly higher in R1 than in R4 ($p<0.001$). There was a non-significant difference in the D/R ratio mismatch between R1 and R2 as well as a non-significant different between R3 and R4 ($p>0.05$).

It was found in R3 and R4 that the mean scaled D/R ratio mismatch is 12 dB greater than the energy ratio found in the BRIR measurements. In the closer positions (R1 and R2), the scaled D/R ratio mismatch is on average 23 dB

greater. The variance component analysis indicated subject-variability is greater in the scaling of D/R ratio than in the judgment of the spatial impression. This indicates larger between-subject variability in terms of expectations regarding matching the acoustical environment that they heard to what they saw on-screen.

In Fig. 10(b), the influence of visual stimuli is shown. While significant, the effect of varying the visual stimulus is less than would be expected given the predictors' leverage with respect to spatial impression. As predicted, when participants saw no sound source in the room, the scaling resulted in a lower D/R ratio mismatch. The D/R ratio mismatch is significantly lower with V1 (no visual stimulus) than with V2 or V3 (loudspeaker, video recording; $p<0.01$). There was no significant difference between D/R ratio mismatch between V1 and V4 (video recording coupled with loudspeakers; $p>0.05$).

In Fig. 10(c), the influence of varying the sound stimuli is shown. In the case of the snare drum sound stimulus (A1), participants scaled those scenes with a significantly lower D/R ratio mismatch than A2, A3, or A5 ($p<0.01$). A non-significant increase was observed in D/R ratio in the judgments of scenes containing A4 (theremin) vs. the A1 sound stimuli ($p>0.05$).

While a significant effect was observed with respect to visual stimuli and sound stimuli, it has been noted in previous research regarding D/R energy-ratio-sensitivity difference limens that human sensitivity is quite coarse. Zahorik³⁶ has reported D/R sensitivity of 5–6 dB while Larsen *et al.* has reported 2–3 dB.³⁷ In the case of the different patterns of means reported in the current investigation, variations in judged D/R ratio expectation may be within this minimum sensitivity region for the visual stimulus and sound stimulus predictors.

The effect noted as a function of receiver position is well outside the difference limen for D/R ratio.

Participants' expectations about the acoustic environment being judged resulted in an overestimation of the direct-to-reverberant energy levels among all tested listening locations, visual stimulus categories, and music or speech performance. This may be due to a conditioning of participants to reproduced music or speech that has been captured with purposely introduced auditory-visual mismatch, namely in the direct-to-reverberant ratio. In many recordings, recording microphones are categorically positioned near the stage, either on tall stands or suspended from cables, in a location much closer than a typical audience seat. In addition to main pickup microphones, recording engineers typically employ several "spot-microphones" in order to increase the clarity of specific instruments or sections of instruments in an ensemble.³⁸

An interesting observation was made during the debriefing process. In response to one post-experiment question regarding source-localization, seven participants indicated that they experienced front-to-back localization confusion in experimental conditions that had no visual presentation accompanying the performance they heard. In contrast, none of the participants indicated that they experienced the same localization confusion when they were presented with an experi-

mental condition which included a visual source stimulus. This could be due to visual capture, where participants will use visual cues to ascertain the location of a sound-emitting source, when available.³⁹ The localization confusion effect for the no-visual-stimulus category can be attributed to the use of non-individualized head-related transfer functions (HRTFs) when recording the BRIR of the testing room (where a dummy-head was used). In future research, the effect can be mitigated by using BRIRs recorded from the ears of the participants in the experiment (through microphones in blocked-ears, for example⁴⁰). While these localization distortions might be present for listeners experiencing a performance with visual cues, the confusion is mitigated by the strong visual feedback that the sound-emitting source is indeed in front of the participant.

Clearly more research is necessary to obtain a complete picture of the importance of providing a coherent visual source stimulus to compliment an acoustic source stimulus while making judgments about the spatial character or quality of an acoustic space. One limitation of the current study was only analyzing the results for main predictor effects. While there were not enough audio-visual scenes to perform a complete full-factorial analysis, preliminary analysis indicated that there were no significant second-order factor interactions in the data set. The inclusion of additional conditions that cover the entire predictor space, including multiple rooms of varying sizes as well as a more complete set of musical instruments are goals for future research. It is clear though, that an interaction exists, and the nature of the visual imagery presented representing a judged sound source can have a significant effect on a participant's expectations in terms of D/R ratio as well as perceived ASW and LEV, corresponding to the participant's spatial impression of a room.

IV. SUMMARY AND CONCLUSIONS

This experiment investigated the influence of visual stimuli variations on participants' scaling of the direct-sound energy and reverberant-energy levels of BRIRs, as well as ASW and LEV judgments. Results showed that the presented visual stimulus variations caused statistically different responses for the same sound stimuli. The presentation of wider-looking sound stimuli (in this case a video recording of an acoustic sound source coupled with two flanking loudspeaker units) increased the perceived ASW and LEV.

The distance to the source as well as the amount of visual cues regarding the size and surfaces of the multi-purpose space also affected a participants' scaling of the spatial parameters. When the subject was presented with a receiver position further away from the source which revealed more of the room's volume, he or she scaled the scene with a greater ASW or LEV. Positive correlations were also found between the measured BRIRs and the scaled values with respect to the BQI and I-IACC for ASW and LEV, respectively. Speech stimuli were scaled with markedly less ASW or LEV for a given scene than was the case for the sound stimuli which included a person playing a musical instrument.

With respect to the D/R ratio mismatch that was collected for each tested audio/visual scene, it was found that in all tested conditions participants adjusted the D/R ratio to greater values than what was found in the original measured BRIRs. The D/R ratio mismatch was further exacerbated in the closer receiver positions (R1 and R2), as opposed to the farther positions (R3 and R4). Significant differences were also found when the visual image of the sound source was exchanged. In those cases, where the video recording of the actual performer or an image of a loudspeaker was presented (conditions V2 and V3), the participant's D/R ratio adjustments resulted in a greater deviation from the actual measured values than was the case for the condition in which an empty room was displayed to the participants (V1). This marked effect of visual sensory information interacting with the expectations of auditory sensory information is important to consider in research that deals with the subjective impressions of room acoustical parameters.

Despite the limited number of conditions tested in this experiment as well as limiting the judged environment to a single multi-purpose space, the results can be generalized to applications where judgments about the quality of the spatial nature of a room are required in a virtual environment. While acousticians understand that the visual nature of a room can have a significant influence on the acoustical impression of that room (Griesinger⁴¹), the influence of the visual sense is generally excluded in experimentation. Currently, it is common practice to judge the quality of a proposed space by listening to a convolved anechoic speech or music sample while looking at a 3-D model of the space without the source being contextualized visually in the space. For convolution with modeled or measured BRIRs, it is common to use pre-recorded source material compiled by external sources such as material from the Archimedes Project.⁴² Very few, if any, high-quality audio-visual recordings of musical instruments or vocal performances in semi-anechoic or anechoic acoustic conditions exist. While it would be impractical to devise a recording of a blue-screened music performance for positioning in a proposed virtual space, similarities in the perception of spaciousness and energy levels were seen between a coherent visual source and surrogate source performance. Positioning a 3-D model of a sound-emitting visual source (i.e., a loudspeaker unit) like the one utilized in this experiment may provide a spatial impression more closely correlated with the spatial perception that would result from having seen the actual visual performance coupled with the presented acoustic source material.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the support for this research from a Humanities, Architecture and Social Sciences (HASS) fellowship granted by Rensselaer Polytechnic Institute. The acoustically treated laboratory space, in which the recordings of the musical instruments and the psychophysical experiments took place, was funded by the New York State Foundation for Science, Technology and Innovation (NYSTAR) as part of the New York STAR Center for Environmental Quality Systems (EQS STAR Center). The

authors also acknowledge Aaron Catlin for the experimental data providing the directivity of the measurement loudspeaker used for this research, and Harisadhan Patra for consulting regarding the statistical analysis. Preparation of this manuscript was supported by T32 Grant No. DC000013. Finally, we thank Masayuki Morimoto for allowing the reproduction of a figure from his previous article¹³ in this article.

- ¹B. E. Stein and M. A. Meredith, *The Merging of the Senses* (MIT, Cambridge, MA, 1993), pp. 3–19.
- ²P. Bertelson and M. Radeau, “Cross-modal bias and perceptual fusion with auditory-visual spatial discordance,” *Percept. Psychophys.* **62**, 321–332 (1981).
- ³W. R. G. Thurlow and C. E. Jack, “Certain determinants of the ‘ventriloquist effect’,” *Percept. Mot. Skills* **36**, 1171–1184 (1973).
- ⁴W. R. G. Thurlow and T. M. Rosenthal, “Further study of existence regions for the ‘ventriloquism effect’,” *J. Am. Audiol. Soc.* **1**, 280–286 (1976).
- ⁵J. Beerends and F. Caluwe, “The influence of video quality on perceived audio quality and vice versa,” *J. Audio Eng. Soc.* **47**, 355–362 (1999).
- ⁶H. McGurk and J. MacDonald, “Hearing lips and seeing voices,” *Nature (London)* **264**, 746–748 (1976).
- ⁷A. H. Marshall, “A note on the importance of room cross-section in concert halls,” *J. Sound Vib.* **5**, 100–112 (1967).
- ⁸M. Barron, “The subjective effects of first reflections in concert halls—The need for lateral reflections,” *J. Sound Vib.* **15**, 475–494 (1971).
- ⁹M. Barron and A. H. Marshall, “Spatial impression due to early lateral reflections in concert halls: The derivation of a physical measure,” *J. Sound Vib.* **77**, 211–232 (1981).
- ¹⁰M. Morimoto and Z. Maekawa, “Auditory spaciousness and envelopment,” in *Proceedings of the 13th International Conference on Acoustics (ICA)*, Belgrade (1989), Vol. **2**, pp. 215–218.
- ¹¹J. S. Bradley and G. A. Soulodre, “Objective measures of listener envelopment,” *J. Acoust. Soc. Am.* **98**, 2590–2597 (1995).
- ¹²J. S. Bradley and G. A. Soulodre, “The influence of late arriving energy on spatial impression,” *J. Acoust. Soc. Am.* **97**, 2263–2271 (1995).
- ¹³M. Morimoto, “The role of reflections from behind the listener in spatial impression,” *J. Applied Acoustics* **62**, 109–124 (2001).
- ¹⁴F. Rumsey, “Spatial quality evaluation for reproduced sound: Terminology, meaning, and a scene-based paradigm,” *J. Audio Eng. Soc.* **50**, 651–666 (2002).
- ¹⁵J. Merimaa and W. Hess, “Training of listeners for evaluation of spatial attributes of sound,” in *Proceedings of the 117th Audio Engineering Society (AES) Convention*, San Francisco, CA (2004), Preprint No. 6237, pp. 1–18.
- ¹⁶J. Blauert and W. Lindemann, “Auditory spaciousness: Some further psychoacoustic analyses,” *J. Acoust. Soc. Am.* **80**, 533–542 (1986).
- ¹⁷P. Larsson, D. D. Västfjäll, and M. Kleiner, “Auditory-visual interaction in real and virtual rooms,” in *Proceedings of the Forum Acusticum, 3rd EAA European Congress on Acoustics*, Sevilla, Spain (2002), Paper No. PSY05-005-IP.
- ¹⁸D. Cabrera, A. Nguyen, and Y. J. Choi, “Auditory versus visual spatial impression: A study of two auditoria,” in *Proceedings of the 10th International Conference on Auditory Display (ICAD)*, Sydney, Australia (2004), pp. 235–242.
- ¹⁹J. Y. Jeon, Y. H. Kim, D. Cabrera, and J. Bassett, “The effect of visual and auditory cues on seat preference in an opera theater,” *J. Acoust. Soc. Am.* **123**, 4272–4282 (2008).
- ²⁰M. Barron, “Subjective study of British concert halls,” *Acustica* **66**, 1–14 (1988).
- ²¹D. Valente and J. Braasch, “Subjective expectations adjustments of early-to-late reverberant energy ratio and reverberation time to match environmental cues of a musical performance,” *Acta. Acust. Acust.* **94**, 840–855 (2008).
- ²²A. McCreery and P. T. Calamia, “Cross modal perception of room acoustics,” *J. Acoust. Soc. Am.* **120**, 3150 (2006).
- ²³A. Farina, “Simultaneous measurement of impulse response and distortion with a swept-sine technique,” 108th AES Convention, Paris, France (2000), Preprint No. 5093 (D-4).
- ²⁴T. Hidaka, L. L. Beranek, and T. Okano, “Interaural cross-correlation, lateral fraction, and low- and high-frequency sound levels as measures of acoustical quality in concert halls,” *J. Acoust. Soc. Am.* **98**, 988–1007 (1995).
- ²⁵“Measurement of the reverberation time of rooms with reference to other acoustical parameters,” International Organization for Standardization (ISO), ISO 3382-Acoustics, 1997.
- ²⁶A. R. Smith and J. F. Blinn, “Blue screen matting,” in *Proceedings of the 1996 SIGGRAPH Annual Conference Series*, 5th–9th 1996, ACM SIGGRAPH, New Orleans, LA (1996), pp. 259–268.
- ²⁷IEEE, “IEEE recommended practice for speech quality measurements,” *IEEE Trans. Audio Electroacoust.* **17**, 225–246 (1969).
- ²⁸Max/Msp/Jitter, <http://cycling74.com/products/max5> (Last viewed 6/6/2009).
- ²⁹J. Sall, A. Lehman, and L. Creighton, *JMP Start Statistics* (Duxbury, Pacific Grove, CA, 2001), Chap. 1.
- ³⁰R. Wolfinger, R. Tobias, and J. Sall, “Computing Gaussian likelihoods and their derivatives for general linear mixed models,” *J. Sci. Comput.* **15**, 1294–1310 (1994).
- ³¹M. G. Kenward and J. H. Roger, “Small sample inference for fixed effects from restricted maximum likelihood,” *Biometrics* **53**, 983–997 (1997).
- ³²J. Becker and M. Sapp, “Synthetic soundfields for the rating of spatial perceptions,” *J. Applied Acoustics* **62**, 217–228 (2001).
- ³³L. Beranek, *Concert Halls and Opera Houses* (Springer-Verlag, New York, 2004), pp. 519–520.
- ³⁴C. A. Brown and W. Yost, “Distance judgments in a reverberant room,” *J. Assoc. Res. Otolaryngol.* **25**, 181 (2002).
- ³⁵C. Nathanail, C. Lavandier, J. D. Polack, and O. Warusfel, “Influence of sensory interactions between vision and audition on the perceptual characterisation of room acoustics,” in *Proceedings of the 1997 International Computer Music Conference (ICMC)*, Thessaloniki, Greece (1997), pp. 414–417.
- ³⁶P. Zahorik, “Direct-to-reverberant energy ratio sensitivity,” *J. Acoust. Soc. Am.* **112**, 2110–2117 (2002).
- ³⁷E. Larsen, N. Iyer, C. R. Lansing, and A. S. Feng, “On the minimum audible difference in direct-to-reverberant energy ratio,” *J. Acoust. Soc. Am.* **124**, 450–461 (2008).
- ³⁸S. R. Alten, *Audio in Media*, 6th ed. (Wadsworth/Thompson Learning, Stamford, CT, 2001), pp. 336–340.
- ³⁹A. Kohlrausch and S. Van De Par, “Audio-visual interaction in the context of multi-media applications,” in *Communication Acoustics (Signals and Communication Technology)*, edited by J. Blauert (Springer, New York, 2005), pp. 109–138.
- ⁴⁰H. Møller, M. F. Sørensen, D. Hammershøj, and C. B. Jensen, “Head-related transfer functions of human subjects,” *J. Audio Eng. Soc.* **43**, 300–321 (1995).
- ⁴¹D. Griesinger, “The psychoacoustics of apparent source width, spaciousness, and envelopment in performance spacing,” *Acust. Acta Acust.* **83**, 721–731 (1997).
- ⁴²V. Hansen and G. Munch, “Making recordings for simulation tests in the Archimedes Project,” *J. Audio Eng. Soc.* **39**, 768–774 (1991).