Comparing perceived auditory width to the visual image of a performing ensemble in contrasting bi-modal environments[a)]

Daniel L. Valente,[b)] Jonas Braasch, and Shane A. Myrbeck[c)]
CA³RL (Communication Acoustics and Aural Architectural Research Laboratory), School of Architecture, Rensselaer Polytechnic Institute, 110 8th Street, Troy, New York 12180

(Received 28 March 2011; revised 13 October 2011; accepted 17 October 2011)

Despite many studies investigating auditory spatial impressions in rooms, few have addressed the impact of simultaneous visual cues on localization and the perception of spaciousness. The current research presents an immersive audiovisual environment in which participants were instructed to make auditory width judgments in dynamic bi-modal settings. The results of these psychophysical tests suggest the importance of congruent audio visual presentation to the ecological interpretation of an auditory scene. Supporting data were accumulated in five rooms of ascending volumes and varying reverberation times. Participants were given an audiovisual matching test in which they were instructed to pan the auditory width of a performing ensemble to a varying set of audio and visual cues in rooms. Results show that both auditory and visual factors affect the collected responses and that the two sensory modalities coincide in distinct interactions. The greatest differences between the panned audio stimuli given a fixed visual width were found in the physical space with the largest volume and the greatest source distance. These results suggest, in this specific instance, a predominance of auditory cues in the spatial analysis of the bi-modal scene.
 [DOI: 10.1121/1.3662055]

PACS number(s): 43.55.Hy, 43.55.Lb, 43.55.Ka [LMW] Pages: 205–217

I. INTRODUCTION

It is commonly understood that visual cues can influence or even dominate acoustic cues during the perceptual analysis of a multi-modal scene. An example of this 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. In contrast, very little is known about how visual cues influence auditory events with respect to other parameters such as spatial impression or how spatial impression generated through the interaction between auditory and visual cues can affect bi-modal perception, which is the subject of the study presented here.

The concept of subjective spatial impression in an enclosure has been explored and defined thoroughly in previous audio-only studies.[5–13] Early investigations by Marshall,[5] Barron,[6] and Barron and Marshall[7] focused on the influence of early lateral room reflections. Later, Bradley and Soulodre[9,10] were able to classify spatial impression into two perceptual components, the apparent source width (ASW) and listener envelopment (LEV).

ASW is defined as 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. Specifically, early arriving lateral reflections (measured within a window of 80 ms after the onset of the direct sound) have been found to cause an increase of ASW.[10]

As the impetus for this study, two studies have investigated the influence of visual cues on the auditory perception of rooms using photos taken in large auditoria.[14,15] Both studies found a strong influence of visual cues on the judgment of auditory events even though the sound-emitting source itself was not included in the visual display. The studies found that subjective impressions of spatial acoustic parameters were statistically different when the participant was presented with a unimodal stimulus (either auditory or visual) as opposed to a bi-modal stimulus (auditory and visual). The importance of including a visual representation of the performance when making acoustic judgments is apparent; it may be difficult to fully measure the subjective expectation with regard to a performer in a space without visually representing the performer in the environment.

While the existence of audiovisual interactions was determined in the aforementioned studies, it remained uncertain to what extent the subjective expectation of a room would change if video footage of the performing musician or talker was included in the visual environment. This question has been addressed in recent work by McCreery and Calamia[16] as well as Valente and Braasch.[17,18] They have examined the effect on the scaling of reverberation time and the direct-to-reverberant energy ratio[16,17] as well as spatial impression[18] with the inclusion of congruent audiovisual stimuli within a judged acoustical environment.

Participants in McCreery and Calamia's experiment[16] exhibited clear expectations of the variance of the direct-to-reverberant ratio with source position based on visual cues.

[a)]A portion of this work was presented at the 2009 meeting of the Audio Engineering Society.
[b)]Author to whom correspondence should be addressed. Electronic mail: daniel.valente@boystown.org. Present address: Center for Hearing Research, Boys Town National Research Hospital, Omaha, NE 68131.
[c)]Present address: Arup, San Francisco, CA 94105.

In Valente and Braasch,[17] it was also shown that participants had a clear expectation regarding the reverberation time of a performance/concert venue once they saw the location and that the expectations regarding reverberation time did not always match results from measurements taken within the acoustic space. Finally, it was found that the visual characteristics of the sound-emitting source affected a subject's expectation regarding spatial impression over and above the acoustic cues provided by the sound stimuli. This was shown with the stimuli being convolved with a simulated room impulse response (IR)[17] or measured binaural room impulse response (BRIR).[18] In both cases, spatial impression was judged differently depending on the visual representation of the sound-emitting source.

While these studies added to the body of knowledge on audiovisual integration, they also posed a number of new questions, some of which are addressed in this paper. In the current study, three questions are addressed. (1) How do participants pan the auditory width of a performing ensemble relative to its geometric visual width in rooms of contrasting reverberation times and physical volumes? (2) How does the physical source-to-receiver distance affect the participants' performance of the task, specifically in highly reverberant environments? Finally (3) does the audio panning of a stimulus conform to geometric cues inherent in the visual width of the sound-emitting source or is it influenced by an increased ASW that the room may more strongly provide through early laterally arriving energy for example?

Depending on the quantity and quality of early arriving lateral reflections, it is well known that the ASW can extend beyond the physical confines of the sound-emitting source. Morimoto and Maekawa[13] describe ASW as the "width of a sound image fused temporally and spatially with the direct sound image" and allude to the fact that this width can be larger than the physical width of the sound-emitting source due to the presence of laterally arriving reflections combining perceptually with the direct sound energy. It is of interest to quantify the difference between the panned acoustic width of a sound source in varying environments and source distances compared to the geometric visual width.

This paper also investigates the extent to which spatial impression varies in a multiple-source scenario. In the authors' previous studies, only solo instruments were used (concert harp and drums,[17] solo violin, djembe, snare drum, theremin, and spoken voice[18]). A larger number of simultaneous musicians were tested in this new study: A traditional vocal quartet consisting of soprano, alto, tenor, and bass vocalists. This ensemble was chosen as a sound stimulus because its geometric width can be parametrically adjusted with ecologically validity; musicians could be spread out to increase their width in a way that would not be possible for a solo instrument.

The paper is organized as follows: Sec. II presents the experimental design describing how the audio-visual stimuli were created and which parameters were under the subject's control during the experiment as well as calibration of the audio-visual system. Section III presents the results of the experiment, and Sec. IV discusses the results as they relate to previous work regarding audio-visual interactions in rooms, as well as exploring subjective impression.

II. METHODS

A. Experimental design

In the investigation, each member of a vocal quartet was recorded on a separate audio channel with an individual video recording. They were recorded such that the acoustic and visual width could be varied independently. It was important to define the angular measures used in this experiment because the study dealt with angular measures in both the auditory and visual domains. Within the context of the current experiment, the term "panning" refers to the lateral spreading of the four audio sources rather than traditional amplitude panning.

During the experiment, the visual imagery of the recorded stimuli was combined with the background image of a visual test room for different locations using monochromatic matting as has been previously done in work by Valente and Braasch.[17,18] For this experiment, the distance between each member of the vocal quartet was adjusted to create the presented visual cues. The physical width of the ensemble before compositing the raw video footage into a 3-D visual model is *ensemble spacing*. This measure is a linear distance.

The angular width of the ensemble as it is seen visually from the perspective of the subject, after being composited into a 3-D room model, is the *visual angular spread*. The visual angular spread is a combination of the ensemble spacing, source distance from the subject, and the physical space that the ensemble is composited into. This is the presented angular width of the ensemble that subjects matched to their perceived *ensemble width* (the lateral audio spreading). This is an angular measure in degrees.

The auditory display featured a hybrid method: The direct sound and early reflections of the ensemble were presented on a 16-channel linear array of loudspeakers positioned behind a projection screen, while the late reverberation was presented across an 8-channel surround array. This reverberation was simulated by real-time convolution of the direct sound of the vocalists with room impulse responses (RIR) created in acoustical modeling software (CATT-ACOUSTIC[21]).

In the experiment, the participants were asked to set the ensemble width of the stimuli to meet their expectations after seeing the visual angular spread of the ensemble. Subjects saw 2-D video recordings of musicians visually composited into 3-D room renderings of rooms reproduced on a 2-D projection screen. A sample frame of the composite video is seen in Fig. 1. Other important parameters such as reverberation time and initial time delay gap were determined to be outside of the aim of this study and thus fixed to the presented visual environment as determined by the virtual acoustical room models. A description of the hybrid auditory display and visual playback system is seen in Fig. 2. This figure shows how the 4-channel recording of the ensemble is split to two computers. The first computer renders the direct sound and early reflections as well as the panning of

FIG. 1. (Color online) A sample frame of video showing a 2-D video recording of a vocal ensemble composited into a 3-D model of an auditorium.

each individual sound source (presented across loudspeaker channels 1–16). The width of the panned audio sources, or ensemble width, is controlled by the subject. The other split stream of audio channels is sent to a second computer that convolves the dry-recorded audio with pre-made impulse responses (IRs) created without the direct-sound component, reproduced over a circular array of loudspeakers (loudspeaker channels 17–24).

The acoustic rendering of five performance spaces (labeled L1–L5) were coupled with 3-D visual renderings of each space at two visual listening positions (a location that was a fixed distance in each space or one that varied by the volume of the space). Imagery of these 3-D models can be seen in Fig. 3. The video recordings of the judged stimuli (the singers) were positioned in each 3-D room model location at a fixed depth with varying ensemble spacing (narrow spacing, NS: 1.5 m; mid spacing, MS: 2.25 m; or wide spacing, WS: 3.0 m). The digital compositing process was the same as used in the authors' previous studies.[17,18]

In the experiment, each participant was presented with a set of experimental conditions that varied the presented stimuli. For this, the visual angular spread of the performance (type of performance space, the location within the judged acoustic space, and the ensemble spacing of the singers) was examined. Subjects were instructed to match their perceived ensemble width to the visual angular spread.

B. Participants

Ten participants, ages 23-36 yr (mean age: 27.4 yr, median age: 25.5 yr), were involved in the experiment. Participants indicated that they had no hearing or uncorrected visual disorder that would prevent them from making the judgments that they would be asked to perform during the test. Participants were questioned regarding their experiences in the fields of audiovisual production, acoustical design, and musical performance.

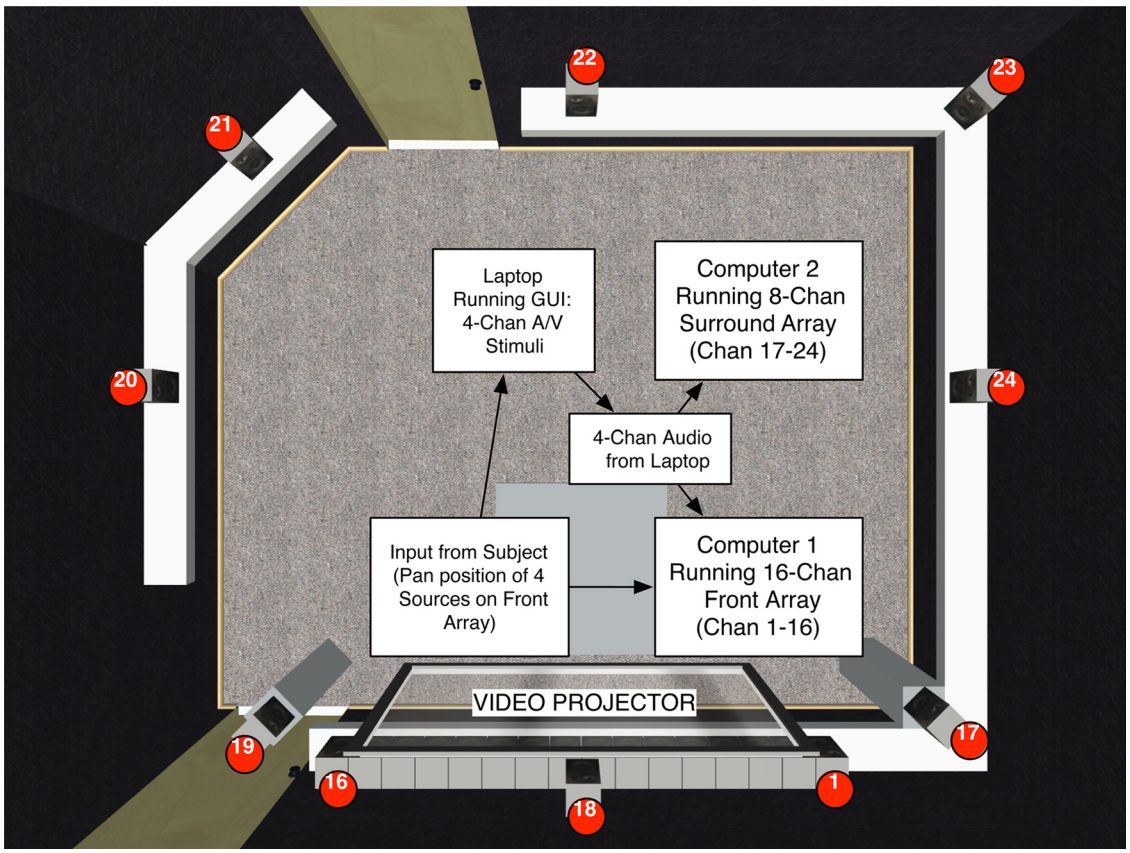

FIG. 2. (Color online) The experimental setup showing both arrays of loudspeakers used to present the direct sound and early reflections of the vocal ensemble (channels 1–16) and the late reverberation of the room model (channels 17–24).

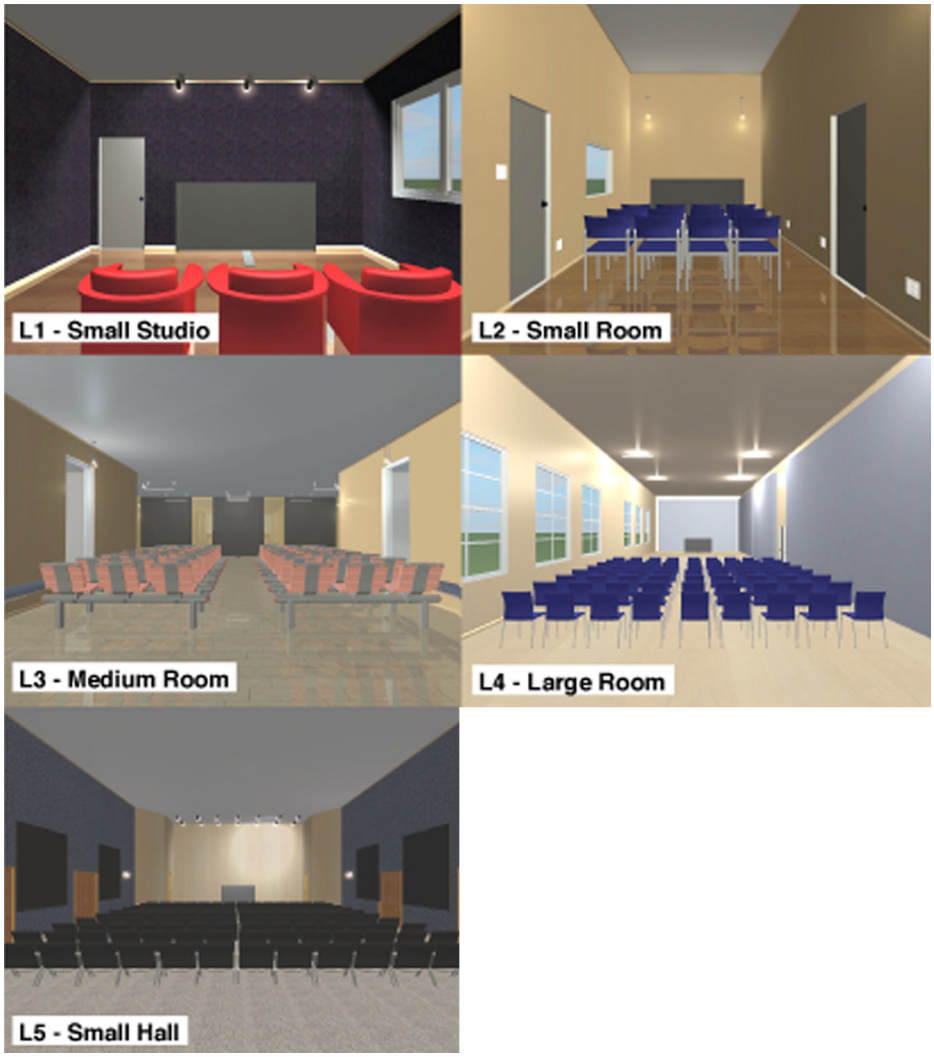

FIG. 3. (Color online) The five physical spaces explored in the study. L1 is a small studio; L2, a small room; L3, a medium-sized room; L4, a large room; and L5, a small hall.

Because this experiment required trained listeners to make detailed audiovisual scaling measures, it was of interest to utilize experienced participants. Musical experience was ranked on the following scale qualified by training in Western Classic music: 1, none (0 years); 2, casual (0 years formal training); 3, trained casual (less than 10 years formal training, not at university level); 4, advanced trained (10 years or more formal training, not at university level); and 5, advanced trained higher degree/professional (one or more college degrees in music, and/or employment by a professional ensemble). Technical experience in acoustics or signal processing was ranked on the following scale: 1, inexperienced (less than 1 year); 2, average (1–3 years of experience); and 3, experienced (3 or more years of experience). The participants' attributes are summarized in Table I.

C. Stimuli

The recorded stimuli were a group of four choral singers arranged in traditional soprano, alto, tenor, and bass configuration. The 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, $3 \times 4 \times 5$ m^3). 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 II. The ambient noise level in the studio was measured at a 15 min $L_{Aeq} < 30$ dB. By using close-microphone techniques and recording the performances in a room with a very short reverberation time (160 ms at 1 kHz), the aim was to minimize room effects such that the

TABLE I. Participants' profiles for the experiment.

Participant	Age	Normal hearing (y/n)	Normal visual (y/n)	Musical experience	Technical experience
1	26	y	y	2	3
2	36	y	y	5	2
3	23	y	y	4	3
4	25	y	y	1	3
5	33	y	y	3	3
6	23	y	y	3	1
7	24	y	y	4	3
8	23	y	y	4	3
9	35	y	y	5	3
10	25	y	y	4	3

TABLE II. 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
C_{50} (dB)	19.54	22.08	28.96	22.61	26.8	25.19
C_{80} (dB)	28.81	32.04	38.91	30.36	34.77	34.34
EDT (s)	0.14	0.13	0.11	0.04	0.06	0.19
T_{30} (s)	0.18	0.17	0.12	0.16	0.15	0.11

recording room would not confound the experiment by adding additional reverberant energy that would not have been present in the virtual room models.

The performers were placed at a distance of 3 m from a blue-screen background and video recorded using an advanced video codec high definition (AVCHD) camcorder (Canon HG10). Monochrome (Chromatte) fabric covered both the area behind and beneath the musicians. To eliminate the need for color-matching in video post-production, tungsten color-temperature lamps were used. Basic three-point (key, fill, and rim) lighting was employed in the studio to achieve an even, natural light quality and good separation of the foreground subject from the background.

To reduce minimal but present audio reflections and to ensure separation for source location adjustments, each singer's performance was recorded on an individual audio channel using a dedicated lavaliere microphone (Sennheiser FP12). A low-noise microphone preamplifier and digital-audio converter was used for the audio-stimulus acquisition (RME Fireface 800). The stimuli were recorded into a digital-audio workstation (DAW) software package (LOGIC PRO 8), and A-weighted sound pressure level (SPL) measurements were taken at a 1 m distance from the source using a type 2 SPL meter (Ono Sokki LA-4240). These SPLs were later used to calibrate the auditory display so that the performances would be presented to participants at the same levels that were measured during the recording session. At the end of the recording session, the musicians were presented with both the auditory and visual performances. Each of them verified that their performance sounded and looked natural to them.

For the experiment, the visual stimuli were reproduced on a 2-D projection screen using a projector capable of life-size reproduction of the recorded stimuli at the reference listening position. The video was presented at the native resolution of the projector: 1280×1024 pixels, exceeding the minimum requirement for high-definition, 720 p video.

The direct sound and first-order early reflections of each singer's audio channel were reproduced by an array of 16 loudspeakers (Yamaha, MSP-5) sized to encompass the width of the projection screen. The auditory display was driven through custom-designed software that simulates an array of microphones positioned in a virtual 3-D recording room. This system is known as *virtual microphone control* (ViMiC) and is a real-time room simulation environment that has been described in detail by Braasch *et al.*[19] ViMiC is based on the concept of positioning an array of virtual microphones and sound sources in a virtual 3-D recording room. The ViMiC virtual environment (which has been released as an open-source module for the real-time audio/visual programming environment: MAX/MSP as part of the Jamoma package[20] is a computer-generated virtual environment that can calculate a gain and delay function from a virtual sound source and virtual microphone receiver. This gain and delay is dependent on the source-receiver distance, axial orientation, and microphone directivity pattern in the same way as an actual sound source is captured by a real microphone. In addition to this direct sound component, early-reflections, as well as a late reverberant tail, can be simulated based on the characteristic surfaces of the virtual recording room.

The same speaker and virtual room modeling setup described by Valente *et al.*[17] was used in the current study with the following adjustments. (1) A linear array of virtual microphones was used as opposed to a circular array, (2) only the direct sound and early reflection modules of the ViMiC software were used, as late reverberation was created separately using convolved IRs on a second system, and (3) subjects were able to control the location of the virtual sound sources (the sound source panning task was the purpose of the experiment).

As such, the virtual sources (the singers, one virtual source per singer) could be positioned at any geometric location within the ViMiC-modeled room relative to a fixed array of virtual microphones based on the physical space and source distance for a given audiovisual scene. Each microphone element in the 16-channel virtual microphone array would pick up each individual singer's direct sound and early reflections and reproduce the appropriately spatialized signal to the subjects.

The ViMiC software was chosen for the experiment as opposed to other techniques such as amplitude-panning due to the desire to reproduce distance cues; ViMiC preserves both amplitude and delay cues to each virtual microphone as well as the simulation of spectral filtering due to sound source distance (high-frequency filtering due to thermal relaxation). In addition, the system operates in real time allowing subjects to adjust the position of the sound sources in the virtual recording room in a way that would not be possible using other methods such as convolution with pre-computed IRs.

The hybrid system used for this study preserves a balance between virtual model-resolution and real-time availability. For this, the direct sound and early reflections (which were required to be panned in real time) could be adjusted during the experiment by the subject, and late reverberation could still be afforded a high-resolution computer model (where the sound sources could be convolved with a single IR independent of panned-sound source position).

The virtual microphone array allowed the subject-positioned ensemble width, which was continuously variable from 0 m width (all virtual sound sources in the same physical location in the center of the array of virtual microphones positioned some distance away as necessitated by the currently assessed visual scene), to 3.5 m (the width of the virtual microphone array) and (real) speakers in the experimental studio.

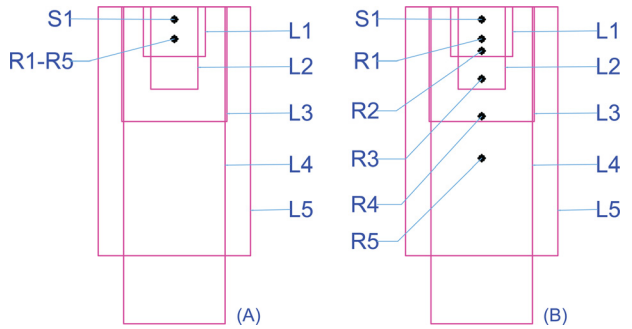

FIG. 4. (Color online) Source (S1) and Receiver (R1–R5) locations with (a) fixed distance and (b) distance varying with the volume of physical spaces L1–L5.

D. Performance rooms

Five audiovisual spaces were created for the experiment. Rather than using photographs and measured impulse responses of existing spaces as has been done in previous research,[14] this experiment created simulated bi-modal environments. The use of simulations prevented participants' preference for or against a real space and also ensured that subjects would not be distracted by any expectations regarding the sound field from a known environment.

Visual models were created with 3-D design software (LIVE INTERIOR 3-D) to have control over lighting, surface composition, and the visual orientation of the room. As in the recording session for the stimuli, tungsten color-temperature lamps with basic three-point lighting were used. The five room simulations were based on existing rooms the textures and visual characteristics of which were mapped onto the renderings based on a site survey. Acoustically, the dimensions of the renderings were adapted to models created in CATT ACOUSTIC,[21] which enables the creation of simulated IRs for real-time convolution with the sound stimuli. These IRs were created without the direct-sound component, as that was rendered by the frontal, 16-channel array. For each room, IRs were created for receivers at two listener positions—one at a fixed distance from the source and another that varied with room volume (see Fig. 4). The fixed distance (3.8 m) represented the widest field of vision based on the farthest possible distance from the source in the smallest room, thus enabling use of the maximum possible variance in the width of the sources. The second position, which varied from room to room as a function of the volume of the room, is given by Eq. (1), where LX_{ds} is the source distance of the physical space x in meters, and LX_{vol} is the volume (m^3) of the physical space: X. This equation was used to create a plausible listening distance for each space, and was based on the maximum possible distance in physical space L1:

$$LX_{ds} = 4.05 + 0.007(LX_{vol}). \quad (1)$$

E. Procedure

Each experimental session consisted of three complete repetitions of the full-factorial treatment combinations of the factors described in Table III. The dependent variables were as follows: *Physical space* (5-level) indicates the virtual audiovisual space in which the choral ensemble was located, *ensemble spacing* (3-level) is the physical width of the ensemble before compositing the raw footage into a 3-D model, and *distance-to-source* (2-level) corresponds to the distance between the judged ensemble and the participant. The distance-to-source was developed to vary between a fixed distance from the participant to the judged stimulus and a distance that varied based on the volume of the physical space in which the performance was presented as described in Eq. (1).

During the experiment, the participant was seated in the center of a dark-walled, sound-treated studio (the same studio used to record to sound stimuli, acoustical parameters found in Table II). The participants were asked to adjust the ensemble width of the four virtual sound sources in the ViMiC environment, laterally controlling the auditory width of the ensemble using a continuous jog wheel until the acoustic rendering aligned with the participant's

TABLE III. Dependent variables for the experiment. The label of *physical space* indicates the virtual space in which the choral ensemble was located, *ensemble spacing* indicates the width at which the choral ensemble was presented, and *distance-to-source* corresponds to the distance metric between the judged ensemble and the participant. Finally the derived visual angular spread is the resultant angle of the choral ensemble given the physical space, distance-to-source, and ensemble spacing.

Condition number	Physical space	Distance-to-source	Ensemble spacing	Visual angular spread (°)
1	L1	Fixed	NS	53.4
2	L1	Volume	NS	53.4
3	L1	Fixed	MS	70.6
4	L1	Volume	MS	70.6
5	L1	Fixed	WS	87.5
6	L1	Volume	WS	87.5
7	L2	Fixed	NS	53.4
8	L2	Volume	NS	44.2
9	L2	Fixed	MS	70.6
10	L2	Volume	MS	56.8
11	L2	Fixed	WS	87.5
12	L2	Volume	WS	71.4
13	L3	Fixed	NS	53.4
14	L3	Volume	NS	28.1
15	L3	Fixed	MS	70.6
16	L3	Volume	MS	38.0
17	L3	Fixed	WS	87.5
18	L3	Volume	WS	46.2
19	L4	Fixed	NS	53.4
20	L4	Volume	NS	19.0
21	L4	Fixed	MS	70.6
22	L4	Volume	MS	24.8
23	L4	Fixed	WS	87.5
24	L4	Volume	WS	31.4
25	L5	Fixed	NS	53.4
26	L5	Volume	NS	16.6
27	L5	Fixed	MS	70.6
28	L5	Volume	MS	23.5
29	L5	Fixed	WS	87.5
30	L5	Volume	WS	29.2

FIG. 5. (Color online) A sample video frame from the calibration stimulus presented.

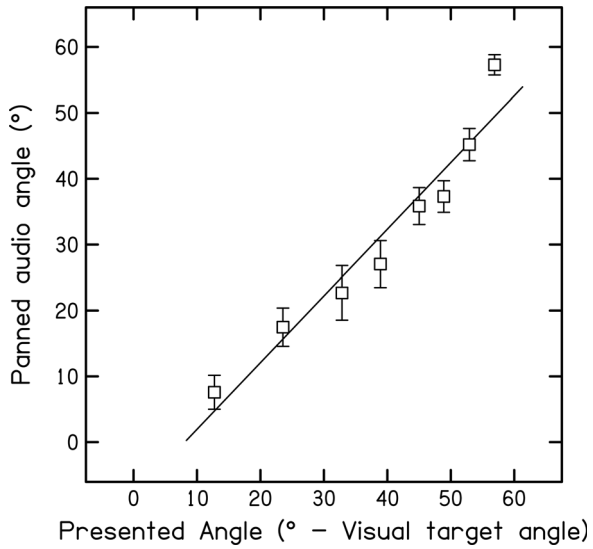

FIG. 6. The audio panning of the sound source is shown on the ordinate versus the visual angle of the target stimuli shown on the abscissa. The mean audio panning angle is shown with ±1 SEM as well as a linear fit of the data.

expectations of how the performance should sound, given the visual angular spread. A jog/shuttle wheel (Griffin Powermate) was used as a physical input to the room-modeling software. It was continuously variable and had the advantage over a linear control-surface fader, for example, in that it did not provide a visual reference of the relative amount of spread within a condition. Turning the shuttle wheel to the left reduced the ensemble width by bringing the location of the four virtual sources closer to the center of the array of virtual microphones in the 3-D virtual recording room within ViMiC. Turning the shuttle wheel to the right increased the ensemble width. The visual angular spread during a trial stayed constant. The task was similar to panning an audio source to the left or right with a set of loudspeakers except an increase in the ensemble width resulted in an equal amount of lateral spread between the four virtual sound sources.

The procedure is outlined next. (1) Participants saw a visual scene of an ensemble positioned at a given source distance in one of the five simulated rooms built from a combination of parameters described in a condition in Table III (the presented visual angular spread). (2) Their task was to control the ensemble width of the acoustic stimuli until that width matched their expectations given the visual scene. (3) Finally, their results were recorded by their indication to proceed to the next scene of the presentation. The sample set was repeated three times to each participant, each time in a new random sequence. For analysis, the mean of three judgments were taken for each subject. Each complete repetition took between 30 and 45 min. Participants were encouraged to take as many breaks as they wished.

Between each trial, the presentation screen was blanked to white to ensure separation between trials. The proceeding instantiation did not begin until activated by the participant. At the conclusion of the experiment, participants were prompted to save their results, distinguished only by their initials, to the desktop of the computer running the experimental software. The results were then renamed with the participant's ID number (generated based on the amount of participants that had completed the test before them) for analysis and presentation.

F. Calibration procedure

An "expert" group of four participants (based on the pre-test survey, a value for 3 for technical experience, 4 or 5 for musical experiment; Table II) were led through a calibration test to establish the ability of a subject to pan a reference acoustic stimulus to a visual target in a reflection-free environment with no room-based cues. For this, a single musician was composited on a black background (see Fig. 5). In this phase, participants were presented with the visual angle of the singer at 5 deg intervals from 0 to 55 deg from the center point of the screen. Each possible visual angle was presented three times with each angular location within each set randomized. A linear regression analysis shows that the subjects' audio panning of the sound source is significantly predicted by the presented visual angle [$F_{(1,30)} = 152.86$, $P < 0.001$, Adj. $R^2 = 0.83$]. This is seen in Fig. 6. Based on this calibration, it was found that the expert group on average underestimated the auditory location of a visual target by an angle of $-8.04°$.

In general, it was found that subjects had greater discrepancies between the presented angle of the visual

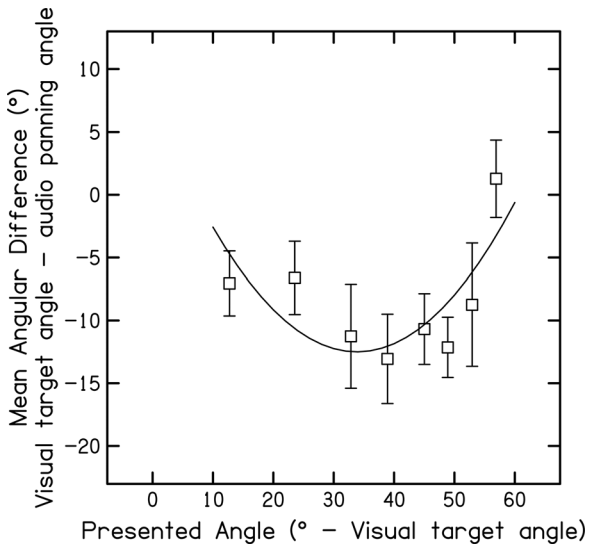

FIG. 7. The mean angular difference between participants' mean audio panning compared to the visual angular position of the calibration stimuli. Mean data are shown with error bars showing ±1 SEM. A second-order polynomial fit to the difference data is shown.

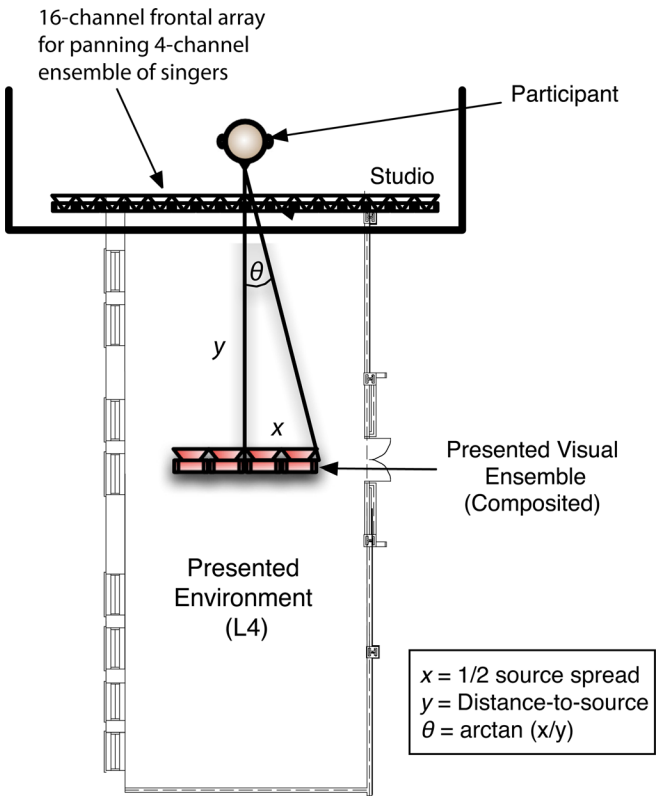

FIG. 8. (Color online) The calculation of the visual angular spread is shown. Based on the distance-to-source metric: Y (either fixed or based on volume, based on volume shown for physical space L4), the visual angular spread of the ensemble is calculated ($2 * \theta$) from the participant's location and the ensemble spacing ($2 * x$).

stimuli and their positioning of the acoustic stimuli in the middle visual angles. This is seen in Fig. 7. This figure shows the mean and standard deviation (SD) for each difference value between auditory and visual stimuli as a function of the visual stimuli. The abscissa of the graph shows the position of the visual stimuli with the ordinate showing the difference between the visual angular location and the subjects' audio panning. This result may indicate edge effects of the task; subjects may have found it easier to pan the corresponding acoustic stimuli to the presented visual target when that target was in the extreme edges of the projection screen. With visual targets in the middle of the screen, subjects likely had less of a reference than when visual targets were near the edge of the screen; this may explain the larger SDs seen in these middle positions. Because subjects set the ensemble width as a whole during the experiment and not as individual sources, this error non-linearity displayed in the calibration phase will become less influential than if subjects were required to set the ensemble width by panning each member of the ensemble individually.

The mean angular difference between the subjects' positioned acoustic stimuli and the visual target shown in Fig. 7 is fit using a second-order polynomial that is a significant predictor of the difference [$F_{(2,29)} = 454.00$, $P = 0.02$, Adj. $R^2 = 0.19$]. This analysis would further support that the error is less toward the physical edges of the screen and greater when subjects were required to pan the audio of a stimulus that was presented near the center of the screen.

Finally, this calibration shows that the auditory display is suitable for presenting the vocal ensemble at their maximum presented visual angular spread of 87.5°. Given these findings, a reductive normalizing adjustment of −8.04° was used to analyze the ensemble width data from each subject for each audiovisual scene tested.

III. RESULTS

A. Calculation of visual ensemble width position influenced by environment cues

The aggregate responses collected from each participant were combined into a sample set for analysis. The sample-set responses were then analyzed to determine any significant interactions between the tested dependent variables and participants' mean responses. Two responses were extracted for the experiment: The raw panned acoustic angle (ensemble width) and a difference angle based on the visual angular spread and the subjects' positioned ensemble width.

For analysis, the specific visual angular spread of the ensemble was calculated based on the presented spacing of the ensemble (factor: Visual source spacing), the physical space, and the distance-to-source (where the based-on-volume parameter level is dependent on the volume of the presented physical space). An explanation of this calculation can be seen in Fig. 8. In this figure, physical space L4 is shown with the source distance based on volume. The participant is shown in the test room looking at a projection of the physical space and the ensemble positioned a distance y meters away. Based on the ensemble spacing, half of the total ensemble spacing is given by a distance x. From these two distances, an angular value θ can be calculated, which doubled gives the visual angular spread.

TABLE IV. Descriptive attributes of the physical spaces used.

Physical space	Volume (m^3)	Distance-to-source fixed (m)	Distance-to-source volume (m)	Reverberation time T_{30} (1 kHz)	IACC fixed position	IACC volume position
L1	67.2	3.8	3.8	0.08	0.51	0.51
L2	84.2	3.8	4.4	0.93	0.84	0.23
L3	204.2	3.8	6.6	0.93	0.49	0.38
L4	841.7	3.8	9.5	2.30	0.71	0.45
L5	1029.7	3.8	11.1	1.34	0.74	0.20

The position of the virtual sound sources in the room model correspond to a distance y meters from the array of virtual microphones (which are positioned virtually at the location of the real speakers in the experimental studio). Subjects control the ensemble width, which spreads the individual sources at a distance x, which may or may not correspond to the visual angular spread of the target stimuli. In the same way as the visual angular spread is calculated, the ensemble width given by a subject's audio panning of the sound stimuli can be expressed as an angle θ given the known y and x coordinates of the audio sound sources.

A description of the physical spaces and derived acoustical parameters of the spaces are shown in Table IV. Table IV shows the volume of each physical space, the distance measurements for each distance-to-source level (the y in Fig. 8), the reverberation time at 1 kHz, and the interaural cross correlation (IACC) at each position derived from the CATT-acoustic model. Here the IACC was calculated from the early component of the simulated impulse response for each space and receiver location. The IACC is reported as a *mid* value, the average of the IACC in the 500, 1000, and 2000-Hz octave bands. In general, the amount of inter-aural coherence is greater in the fixed receiver positions and in physical spaces that have a smaller volume.

A three-way analysis of variance (ANOVA) was fit to subjects' mean data collected for each audiovisual scene. The three main effects (physical space, visual source spacing, and distance-to-source) were explored as well as the interactions between factors. Based on the known visual angular spread of the audiovisual scene (given in Table III) and the collected ensemble width by the participant, a difference factor between the two can be realized, representing the variation between the visual angular spread of the ensemble in a scene and the mean subjects' reported ensemble width. A positive variation indicates that the subjects' positioned ensemble width was larger compared to the visual angular spread.

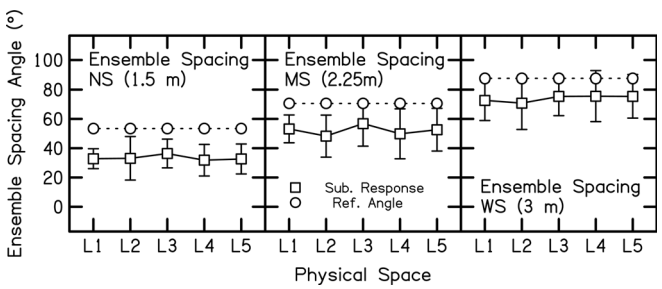

FIG. 9. A comparison of the subject-positioned ensemble width (solid lines) to the known visual angular spread (dashed lines) of the performing ensemble in physical spaces: L1–L5. These data are for the given physical space combined with the three ensemble spacing categories tested. The source distance varies by volume. The ensemble spacing is labeled by NS, MS, and WS. The error bars in the auditory angle are constructed using ±1 SEM.

B. Ensemble width

For the positioned ensemble width for each audiovisual scene, the whole model ANOVA was found to be significant [$F_{(29,270)} = 11.53$, $P < 0.001$, Adj. $R^2 = 0.51$]. While the three-way interaction among physical space, distance-to-source, and ensemble spacing was not found to be significant, significant two-way interactions between ensemble spacing and distance-to-source [$F_{(2,270)} = 8.69$, $P < 0.001$] as well as physical space and distance-to-source [$F_{(4,270)} = 4.89$, $P < 0.01$] were found.

Given the ability of subjects in the calibration phase to position the acoustic stimuli differently given changing target visual stimuli, it is no surprise that different visual angular widths due to the combination of dependent variables of each audiovisual scene caused the subjects' ensemble width of the acoustic stimuli to change. The ensemble width was found to be dependent on the ensemble spacing, distance to source, and physical space. Figure 9 shows the mean across subjects and ±1 SEM of the ensemble width for each of the scenes in the fixed source-distance conditions compared to the visual angular spread of each presented choral ensemble. Figure 10 shows the same comparison for the conditions where the source distances were based on the volume of the physical space. The physical space by ensemble spacing second-order interaction was not found to be significant ($P > 0.05$), indicating the relative impact of the individual ensemble spacing (NS, MS, or WS) on the ensemble width

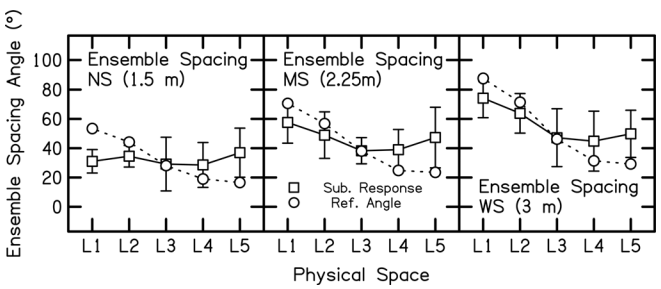

FIG. 10. A comparison of the subject-positioned ensemble width (solid lines) to the known visual angular spread (dashed lines). The error bars in the auditory angle are constructed using ±1 SEM.

judgment was not impacted differently across the physical spaces.

Significant main effects of physical space [$F_{(4,270)} = 2.95$, $P < 0.02$], ensemble spacing [$F_{(2,270)} = 122.27$, $P < 0.001$], and distance-to-source [$F_{(1,270)} = 24.88$, $P < 0.001$] were also found but must be considered in light of the significant interactions. The perceived ensemble width given the presented visual angular spread is dependent on the physical space, ensemble spacing, and distance-to-source but has a different perceived ensemble width depending on the combination of these factors.

This can be seen in Fig. 9, where the visual angular spread is fixed across physical space but varies as a function of ensemble spacing (visual angular spread increases as the ensemble spacing increases from NS to WS). This is also reflected in the subjects' ensemble width data, which stays relatively constant as function of physical space but increases within wider ensemble spacing.

In Fig. 10, the visual angular spread changes based on the physical space (which is now based on the volume of physical space). As a result, the visual angular spread decreases as the volume of the physical space increases. This occurs as a function of source distance. In other words, an ensemble that is further away from the subject is seen as having a narrower visual angular spread despite being positioned with the same ensemble spacing. In this case, the ensemble width data follow the trend of a decreasing angle as a function of increasing physical space volume in the mid-spaced and widely spaced ensembles in physical spaces L1–L3 but does not decrease monotonically. In the case of the narrow-spaced ensemble, the ensemble width positioned by the subject does not follow the trend of decreasing as a function of physical space.

As the choral ensemble's visual angular spread increased, the participants' positioned ensemble width also increased. This is an expected and meaningful effect that speaks to the participants' ability to accomplish the task given to them. Given a wider visual angular spread, participants reacted to the task by positioning the location of the acoustic sources more widely (increasing the ensemble width). Within that effect, though, lies an interesting result, which is the increasing lack of congruency between the visual angular spread of the ensemble and the positioned ensemble width, necessitating the analysis of the angular difference between the visual angular spread and the positioned ensemble width of the acoustic stimuli by subjects.

C. Ensemble width compared to visual angular spread

To further assess the difference between subjects' positioned ensemble width and the visual angular spread for each condition, the difference between the two angular measurements was addressed by a full-factorial ANOVA. The whole model was found to be significant [$F_{(29,270)} = 9.54$, $P < 0.001$, Adj. $R^2 = 0.45$]. The three-way interaction among physical space, source, and ensemble spacing was not significant; however, the physical space by source distance interaction was significant [$F_{(4,270)} = 15.80$, $P < 0.001$]. This interaction can be seen in Fig. 11.

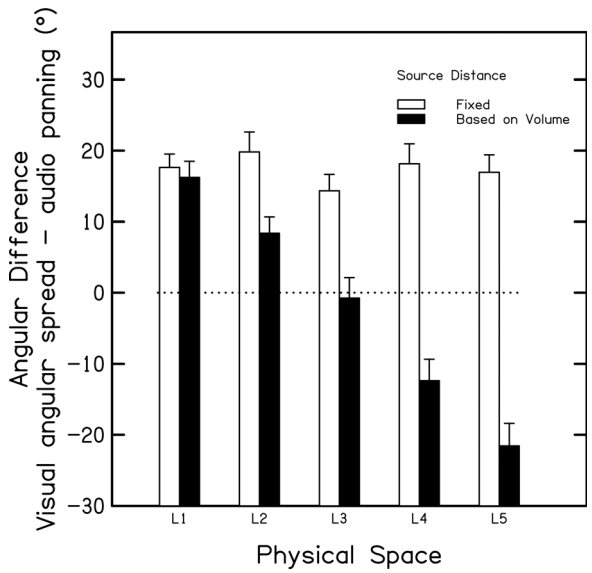

FIG. 11. The angular difference between the mean subjects' positioned ensemble width compared to the visual angular spread across source distance and physical space. The error bars are constructed by ±1 SEM.

On the abscissa of the graph, the physical spaces are shown with the angular difference between the ensemble width and the visual angular spread as a function of source distance. As ensemble spacing was not found to be a significant predictor of this difference metric ($P > 0.05$), the mean values across subjects are collapsed across the three levels of ensemble spacing. The difference is dependent on both the physical space and the source distance. The relative difference in the fixed source distance is similar despite increasing acoustic enclosure volume (and increasing reverberation) but decreases when the source distance is increased based on volume. In addition to the significant interaction, both main effects of physical space [$F_{(4,270)} = 17.67$, $P < 0.001$] and source distance [$F_{(1,270)} = 132.7$, $P < 0.001$] are significant but must be considered in light of the significant interaction between the two predictors.

As the audiovisual scene of the ensemble was presented further away from the participant an interesting interaction was apparent: The angular difference between the visual angular spread and the positioned ensemble width varied as function of the volume of the physical space.

D. Correlation of IACC to normalized auditory width

A significant pair-wise correlation was found between the aforementioned audiovisual difference measure and the inter-aural cross correlation (IACC, see Table IV). The latter represents a measure of the difference value in sounds received by two ears of a listener and effectively collects all of the laterally arriving reflections in the first 80 ms from the direct sound. High values of IACC (close to the maximum possible value of 1.0) represent highly correlated, nearly identical signals at each ear. Values of IACC that are close to 0.0 represent signals that have little or no correlation at all. For more information about IACC, the reader may consult Hidaka *et al.*[22]). Analysis revealed a significant, moderately positive correlation of 0.49 ($P = 0.001$, 95% CI [0.40,

0.57]) between a simulated IR's IACC and the difference value between subjects' positioned ensemble width and the visual angular spread. In listening/viewing locations that have a lower value for IACC, a negative difference value is seen that corresponds to the participants positioning the ensemble width wider than the visual angular spread. In listening/viewing locations that have a higher value of IACC, a positive difference is seen that corresponds to the participants positioning the ensemble width narrower than the visual angular spread.

IV. DISCUSSION

This experiment investigated cues for integration and sensory dominance of audition and vision, in the tradition of Stein and Meredith,[4] among others. The novelty of the current research is that it presented an environment comprised of the stimuli in question rather than, like Vroomen and DeGelder, using merely test tones and simple light sources[23] or isolated auditory and visual stimuli as did Cabrera *et al.*[14] The authors have employed and expanded upon developments by Blauert *et al.*,[24] Larsson *et al.*,[25] Braasch *et al.*,[20] and others in creating an immersive and versatile electroacoustic environment.

The group of participants tested revealed an inverse relationship between the positioning of ensemble width of an auditory stimulus given a visual angular spread and reverberation time. There is precedence for these findings in the work of Bradley and Soulodre[9] as well as Morimoto and Maekawa,[13] who demonstrated the influence of higher degrees of ASW on the sensitivity to early arriving lateral reflections. The results of this experiment expand upon their findings by revealing a relationship between increasing distance-to-source and participants' expectations regarding auditory width given a simultaneous presented visual angular spread of stimuli. In this scenario, the presence of visual cues is unlikely to assist in this, and judgments of audiovisual alignments are shown to deviate from the target visual stimuli (as seen in Fig. 11).

An interesting correlation between the resulting IACC at the listening/viewing position and the difference between audiovisual matching was found. IACC and the derived *binaural quality index* (BQI), which corresponds to 1-IACC, has been shown to be an effective metric for rating the sound quality index in a concert hall, and in general, listeners prefer spaces with a low IACC (non-correlated sounds arriving at both ears of the listener).[26] While the experiment was not focused on rating the sound quality of the assessed audiovisual environments, this correlation is particularly illuminating, as the IACC metric collects all of the differences at the ears from laterally-arriving reflections in the first 80 ms from the direct sound. These lateral reflections have been previously shown to influence ASW.[9] Therefore in the spaces that provided an increased amount of uncorrelated lateral reflections (represented by a lower IACC), participants judged the ensemble to be greater than the visual angular spread. The increased lateral reflections resulted in a larger ASW, and therefore greater ensemble width may have been required to match the visual angular spread of the ensemble that the participants were shown.

In spaces that presented highly correlated (IACC closer to 1.0) sounds at both ears, participants judged the ensemble to be, on average, narrower than the visual angular spread. The highly correlated sounds that arrived at both ears of the participants resulted in a lower ASW, and as a result participants may have compensated by reducing the ensemble width.

It is clear that the influence of various factors in an acoustic environment can affect the difference between the positioned ensemble widths compared to visual angular spread of a stimulus. The spatial impression created by room reflections varies based on the amount of laterally arriving early reflections that reach the ears of the listener/viewer. As a result, recreating an ensemble width based solely on the known visual angular spread of an ensemble will not necessarily align with the expectations of the participant. It seems quite important to take into account the acoustical response of the environment. This is particularly important regarding the density and magnitude of early arriving lateral reflections within 80 ms of the arrival of the direct sound, as they will be the most influential reflections in affecting the ASW of that ensemble or sound. After this is taken into account, one must make adjustments according to how the room itself will influence the expectations of a listener/viewer experiencing a performance located in that acoustical environment to present the ensemble width to match the visual angular spread.

V. CONCLUSION

The experiment that was carried out presented an immersive audiovisual environment in which participants were instructed to set the ensemble width of a vocal quartet given a presented visual angular spread of that ensemble in a dynamic cross-modal simulated room environment. Supporting data were accumulated in two stages: First through a calibration matching test in a reflection-free environment, where "expert" acousticians performed an audiovisual matching test to test the validity of the experimental system and second where participants were given an audiovisual panning test in rooms of increasing physical volumes. In this case, subjects were instructed to align the ensemble width to a varying set of audio and visual cues and visual angular spreads. Both auditory and visual cues were found to affect the collected responses, and the two sensory modalities interacted.

This experiment aimed to address three research questions. It was first of interest to determine how well participants panned the auditory width of a performing ensemble relative to its geometric visual width in rooms of contrasting reverberation time and physical volumes. It was shown that in small rooms with short reverberation times and in large rooms where the source-to-receiver distance was small, participants are able to reliably match the ensemble width to the visual angular spread. The second research question that was posed asked: How does the physical source-to-receiver distance affect the participants' performance of the task,

specifically in highly reverberant environments? When looking at the angular difference between the visual angular spread and the positioned ensemble width, conditions that included a greater source-to-receiver distance resulted in a larger difference between these two measures. The performance of the task as a function of the source-to-receiver distance was not affected in the sense of greater variance across responses but rather a systematic difference between the visual angular spread that was shown and the corresponding positioned ensemble width.

Finally, the question was asked: Does the audio panning of a stimulus conform to geometric cues inherent in the visual width of the sound-emitting source (visual angular spread) or are they influenced by an increased ASW that the room may more strongly provide, through early laterally arriving reflections, for example? This question is addressed in the significant correlation that was found between the IACC and the difference between the positioned ensemble width and the visual angular spread. Spaces that provided an increased amount of uncorrelated lateral reflections, or a lower IACC, resulted in participants positioning the ensemble width greater than the visual angular spread. In spaces that presented highly correlated (IACC closer to 1.0) sounds at both ears, participants positioned the ensemble width, on average, narrower than the visual angular spread. These results suggest, in this specific instance, a predominance of auditory cues in the spatial analysis of the bi-modal scene, in contrast to what has been reported in previous studies (e.g., the ventriloquist effect).[1–3]

The research explored here has provided a technique that accurately presents both auditory and visual cues that occur in an environment and provides interactive control of them in a studio setting for experimental purposes. By providing life-sized video recordings visually composited within models of acoustical environments, the visual presentation scheme used in this research, when coupled with the multi-channel auditory display, is more ecologically valid when compared to previous work that has not included imagery of the sound-emitting source or smaller visual representation (black-and-white photographs, for example) of the judged environment. While the number of subjects used for this study was small, the fact that significant differences were seen across the conditions investigated shows that the auditory-visual interactions presented here are important parameters in the subjective assessment of a space. Nonetheless, a follow-up study with a larger group of subjects should be conducted to verify the significant results presented in the current study as well as to explore several of the factors not found to be significant, such as relative ensemble spacing, to see if they become statistically significant in a larger subject population. In addition, the research should be extended to include a larger number of musical ensembles and styles as well as a more diverse set of acoustical environments.

Finally, it is important that researchers and designers of acoustic spaces consider that the expectations regarding room acoustics parameters and spatial impression of a space are based on both auditory and visual cues. The results of the experiment performed in this paper help to expand the understanding of the complex audiovisual relationship among a room, sound- emitting source, and a person sensing his or her environment.

ACKNOWLEDGMENT

The authors gratefully acknowledge the partial support for this research from a Humanities, Architecture and Social Sciences (HASS) fellowship granted by Rensselaer Polytechnic Institute (RPI). 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). Preparation of this manuscript was supported by NIH T32 DC000013. The authors finally thank the helpful comments and suggestions provided by two anonymous reviewers on an earlier version of this manuscript.

[1]W. R. G. Thurlow and C. E. Jack, "Certain determinants of the 'ventriloquist effect'," J. Percept. Mot. Skills **36**, 1171–1184 (1973).
[2]W. R. G. Thurlow and C. E. Jack, "Further study of existence regions for the 'ventriloquism effect'," J. Am. Audiol. Soc. **1**, 280–286 (1976).
[3]P. Bertelson and M. Radeau, "Cross-modal bias and perceptual fusion with auditory-visual spatial discordance," Percept. Psychophys. **62**, 321–332 (1981).
[4]B. E. Stein and M. A. Meredith, *The Merging of the Sense* (MIT, Cambridge, MA, 1993), pp. 3–19.
[5]A. H. Marshall, "A note on the importance of room cross-section in concert halls," J. Sound Vib. **5**, 100–112 (1967).
[6]M. Barron, "The subjective effects of first reflections in concert halls—the need for lateral reflections," J. Sound Vib. **15**, 475–494 (1971).
[7]M. 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).
[8]J. Blauert and W. Lindemann, "Auditory spaciousness: Some further psychoacoustic analyses," J. Acoust. Soc. Am. **80**, 533–542 (1986).
[9]J. S. Bradley and G. A. Soulodre, "The influence of late arriving energy on spatial impression," J. Acoust. Soc. Am. **97**, 2263–2271 (1995).
[10]J. S. Bradley and G. A. Soulodre, "Objective measures of listener envelopment," J. Acoust. Soc. Am. **98**, 2590–2597 (1995).
[11]F. Rumsey, "Spatial quality evaluation for reproduced sound: Terminology, meaning, and a scene-based paradigm," J. Audio Eng. Soc. **50**, 651–666 (2002).
[12]J. 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.
[13]M. 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.
[14]D. 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.
[15]P. Larsson, D. D. Västfäll, 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.
[16]A. McCreery and P. T. Calamia, "Cross modal perception of room acoustics," J. Acoust. Soc. Am. **120**, 3150 (2006).
[17]D. L. 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).
[18]D. L. Valente and J. Braasch, "Subjective scaling of room acoustical parameters influenced by visual environmental cues," J. Acoust. Soc. Am. **128**, 1952–1964 (2010).

[19]J. Braasch, N. Peters, and D. L. Valente, "A loudspeaker-based projection technique for spatial music applications using virtual microphone control," *C*omput. Music J. **3**, 55–71 (2008).
[20]"Jamoma: A platform for interactive art-based research and performance," from http://www.jamoma.org/ (Last viewed July 22, 2011).
[21]CATT, Mariagatan 16A, SE-41471 Gothenburg, Sweden. http://www.catt.se/ (Last viewed June 6, 2009).
[22]T. Hidaka, 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).
[23]J. Vroomen and B. de Gelder, "Perceptual effects of cross-modal stimulation: Ventriloquism and the freezing phenomenon," in *The Handbook of Multisensory Processes* edited by G. Calvert, C. Spense, and B. E. Stein (MIT, Cambridge, MA, 2004), pp. 141–150.
[24]J. Blauert, H. Lehnert, J. Sahrhage, and H. Strauss, "An interactive virtual-environment generator for psychoacoustic research," Acta. Acust. Acust. **86**, 94–102 (2000).
[25]P. Larsson, D. D. Västfäll, and M. Kleiner, "Ecological acoustics and the multi-modal perception of rooms: Real and unreal experiences of audiovisual virtual environments," in *Proceedings of International Conference on Auditory Display (ICAD),* Helsinki, Finland (2001), pp. 245–249.
[26]L. Beranek, *Concert Halls and Opera Houses* (Springer-Verlag, New York, 2004), pp. 519–520.