Audio Engineering Society

Conference Paper

Presented at the Conference on
Audio for Virtual and Augmented Reality
2018 August 20 – 22, Redmond, WA, USA

Toward objective measures of auditory co-immersion in virtual and augmented reality

G. Christopher Stecker[1,2,3], Travis M. Moore[1], Monica Folkerts[1], Dmitry Zotkin[4], and Ramani Duraiswami[4]

[1] *Vanderbilt University School of Medicine, Nashville TN USA*
[2] *Vanderbilt Brain Institute, Nasvhille TN USA*
[3] *Auditory Space, LLC, Sheridan WY USA*
[4] *Visisonics Corp., College Park MD USA*

Correspondence should be addressed to G. Christopher Stecker (cstecker@spatialhearing.org)

ABSTRACT

"Co-immersion" refers to the perception of real or virtual objects as contained within or belonging to a shared multisensory scene. Environmental features such as lighting and reverberation contribute to the experience of co-immersion even when awareness of those features is not explicit. Objective measures of co-immersion are needed to validate user experience and accessibility in augmented-reality applications, particularly those that aim for "face-to-face" quality. Here, we describe an approach that combines psychophysical measurement with virtual-reality games to assess users' sensitivity to room-acoustic differences across concurrent talkers in a simulated complex scene. Eliminating the need for explicit judgments, Odd-one-out tasks allow psychophysical thresholds to be measured and compared directly across devices, algorithms, and user populations. Supported by NIH-R41-DC16578.

1 Introduction

Auditory spatial awareness (ASA) is the important skill of using sound to understand extrapersonal space and its content. Both explicit (sound localization) and implicit (environmental awareness) aspects of ASA contribute to everyday listening, effort, fatigue, and spatial attention. Because hearing alone provides spatial information in all directions, ASA is critical for awareness outside the visual field and for the experience of sensory "immersion" in both natural and virtual scenes. Studies suggest that hearing impairment, aging, and device use can negatively impact ASA, but the current lack of tools for objective psychophysical assessment in realistic spatial scenes remains a significant barrier to progress in this area. Virtual immersive experiences with accurate loudspeaker-based or binaural room simulations offer new approaches to measuring listeners' sensitivity to auditory spatial information in realistic scenes.

A similar challenge stands in the way of validating immersive experiences themselves. For example, in auditory augmented reality (AR), synthetic or recorded sound overlays the natural sounds of the physical environment. These virtual sound layers may be presented using earphone inserts that also allow natural sound to enter through an acoustic vent, or can be electronically mixed with natural sounds captured by external microphones. AR applications may present sound alone (e.g., spatially aware hearing aids) or in combination with vision (geo-descriptive sound tagging). Importantly, applications that aim to achieve “face-to-face” quality require *co-immersion* of natural and artificial sounds. That is, objects should be perceived as immersed within or belonging to a shared perceptual scene, rather than artificially superimposed upon it, but few measures are available to objectively assess the quality of co-immersion between objects in a scene.

Co-immersion of real objects in natural scenes is enhanced by environmental features such as common lighting and reverberation. Co-immersion in purely artificial scenes (e.g. video games or music mixes) can be similarly enhanced by applying consistent lighting and reverberation across objects, or it may be intentionally reduced with distinct lighting and reverberation effects, for example to better distinguish foreground from background elements. Compellingly realistic AR, however, requires seamless integration of natural and synthetic elements. That goal, in turn, requires that environment features be (1) estimated from the natural elements and (2) appropriately recreated for the synthetic elements. Objective measures of co-immersion are necessary to validate such approaches and quantify differences between algorithms, devices, and user populations (e.g. aging and hearing-impaired users). One possible approach is to measure users’ ability to discriminate a virtual target from natural background sounds, or equivalently to discriminate among synthetic targets treated with different synthesis features (e.g. room acoustics). Strong co-immersion should reduce listeners’ sensitivity to such differences; hence, such discrimination measures could be used to validate synthesis quality and/or to assess ASA among different populations of potential users (e.g. children, aging, or hearing-imparied listeners).

This paper describes one approach to objectively measure co-immersion in virtual auditory scenes. The study used a simple virtual-reality (VR) environment to simulate a complex scene involving several (typically, four) concurrent familiar talkers. Prior to testing, users were trained to recognize each of six talkers by voice and speech topic, and to associate each talker with an assigned name and cartoon face. Co-immersion was quantified by measuring users’ sensitivity to differences in the room acoustics applied to each talker. Users performed a pair of tasks (spatial localization and talker identification) designed to simulate aspects of listening in complex natural scenes.

On each trial, one talker was selected as the target. Audio from the target talker was processed via a room-acoustical model that varied in room shape or surface characteristics from trial to trial. Audio from the other talkers was processed via the same model, but configured for a common standard room which remained fixed from trial to trial. Audio, including direct sound and 13 orders of simulated reflection, was presented via a 360° circular array of 64 loudspeakers in the Vanderbilt Bill Wilkerson Anechoic Chamber Lab (ACL). Users localized and identified the target talker by interacting with the VR display, and received feedback that indicated the correct target location along with the positions and identities of all talkers.

The results indicate reliable discrimination of room acoustics in all three conditions, and were similar for both localization and talker-identification tasks. Unexpectedly, thresholds in the Size+ condition tended to be greater than in reverberation-matched Refl+ conditions, suggesting that judgments did not rely exclusively on reverberation time. More importantly, the results serve as proof-of-concept regarding this approach. Future work could use similar techniques to quantitatively validate co-immersion in auditory AR, and to compare spatialization algorithms in VR-audio workflows. The approach is particularly well suited to assessing questions of universal accessibility, for example by obtaining measures in individuals and in user populations (e.g., children, elderly, or sensory-impaired users) who may likely differ in their capacity to use auditory spatial information.

2 Methods

Participants were six normal-hearing young-adult listeners (four female). Two participants was employed in the lab; others were paid participants naive to the purpose and hypotheses of the study.

Stimuli consisted of short phrases from the Edinburgh University Speech Timing Archive and Corpus of English [1]. The corpus comprises phonetically controlled sentences produced by six talkers (three female). For

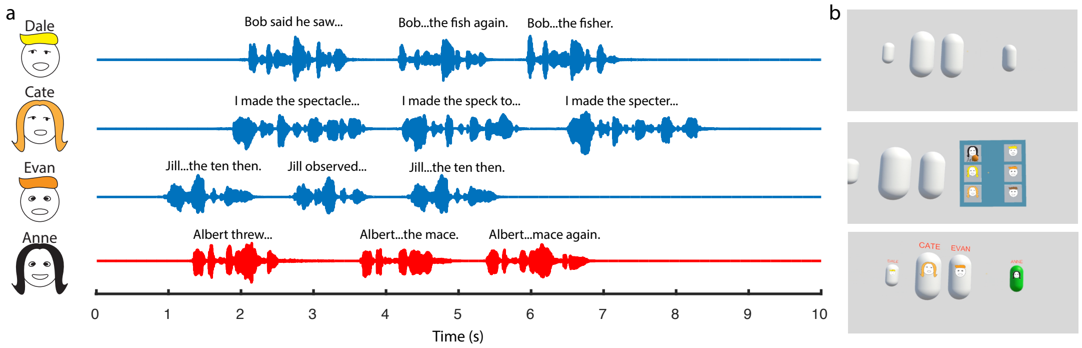

Fig. 1: Timeline (a) and response sequence (b) of a single trial in the main experiment. Four concurrent talkers each produced three sentences from the list; one was designated the target (Anne in this example) and acoustically rendered in a different room model than other talkers. Images (b) simulate the participant's view in the head-mounted display. Blank markers indicated talker locations at the start of each trial. Participants (1) localized by head pointing and (2) identified the target talker from a list. Feedback (3) indicated the target marker location (green if identified correctly, red otherwise) and identities of all talkers.

the current study, we selected a single list of four related sentences for each talker (Table 2). Each talker was paired with a cartoon image and name (see Fig. 1) with which participants learned to identify talkers by voice and sentence list during the training phase. Recordings were normalized in RMS amplitude across sentences and talkers and presented over loudspeakers at 50 dBA SPL per talker.

Stimuli were presented via an array of 64 ear-height loudspeakers spanning 360° of azimuth. The loudspeaker array implemented a virtual room simulation which included the first 13 orders of lateral reflection (front, back, and side walls) in a rectangular room whose dimensions or reflection coefficients varied across talkers. Reflection azimuth, timing, and intensity were computed using the image method as detailed by Stecker and Moore [2]. All but one of the talkers were presented with reflections corresponding to a standard room (Room 0) with dimensions 10m x 10m and surface reflection coefficient of 0.5. The target talker on each trial was presented with reflections corresponding to a different room model. Target rooms varied in either size (larger than standard: condition *Size+*) or surface reflection (greater [condition *Refl+*] or less [*Refl-*] than standard in different conditions). Details of the various rooms are given in Table 1. Note that rooms smaller than standard were omitted to avoid positioning talkers outside the virtual room boundaries. Target room configurations were established during pilot testing to span the full range of chance to perfect discrimination.

In order to reduce the impact of absolute intensity cues on reverberation judgments, auditory stimuli were presented without intensity correction for distance, but with a random level rove of ±5 dB across talkers. Similarly, reflection amplitudes were not adjusted for distance but were instead scaled by the product of surface reflectivity and reflection order to accurately simulate acoustic absorption.

2.1 Apparatus

Testing was conducted in the Vanderbilt Bill Wilkerson Center Anechoic Chamber Laboratory (ACL). The ACL consists of a large (4.6 x 6.4 x 6.7 m) anechoic chamber (Eckel Industries; Cambridge MA USA) and circular array (2m radius) of 64 ear-height loudspeakers spanning 360° azimuth (spacing of 5.625°). Loudspeakers (Meyer MM-4, Berkeley CA USA) were driven by digital amplifiers (Ashly ne8250pe, Webster NY USA) controlled by a dedicated Dante audio-over-ethernet network (Focusrite Rednet, El Segundo CA USA). Experiments were controlled and audio delivered from a Mac Pro workstation (Apple, Cupertino CA USA) running MATLAB 2017b (Mathworks, Natick MA USA).

Visual information was presented using a head-mounted display (HTC Vive, New Taipei City Taiwan) with hand controllers for user interaction. This

	N	Param	ITDG	C50	RT60	LF
Size+	0	10.0 m	16 ms	6.5 dB	.29 s	.53
	1	12.0 m	22 ms	4.9 dB	.36 s	.52
	2	14.4 m	27 ms	2.9 dB	.44 s	.51
	3	17.3 m	34 ms	0.2 dB	.54 s	.52
	4	20.7 m	41 ms	-1.8 dB	.65 s	.50
Refl+	0	0.5	16 ms	6.5 dB	.29 s	.53
	1	0.57	16 ms	3.9 dB	.36 s	.52
	2	0.64	16 ms	1.5 dB	.46 s	.52
	3	0.70	16 ms	-0.6 dB	.55 s	.52
	4	0.74	16 ms	-2.2 dB	.66 s	.52
Refl-	0	0.5	16 ms	6.5 dB	.29 s	.53
	1	0.41	16 ms	10.3 dB	.24 s	.53
	2	0.32	16 ms	14.5 dB	.19 s	.53
	3	0.25	16 ms	18.7 dB	.15 s	.53
	4	0.00	n/a	-inf dB	.00 s	.53

Table 1: Room models tested in each condition. N: room index, increasing with difference from 0 (standard room). Param: manipulated parameter. In condition Size+, this is the room width (=length) in meters. In conditions Refl+ and Refl- this is the surface reflection coefficient ($1-\alpha$). ITDG: initial time-delay gap in ms, the time between arrival of direct sound and first early reflection. C50: ratio, in dB, of energy arriving within 0–50 ms of direct sound, relative to energy arriving after 50 ms. RT60: time (s) for reverberation to decay by 60 dB. LF: lateral energy fraction calculated 5–80 ms after direct sound.

VR apparatus was controlled by a custom PC running Steam VR (version 2017-01-30, Valve Corporation, Bellevue WA USA) and A-SPACE (version 0.1, www.auditory.space), a custom VR software application that can be controlled from MATLAB. In this case, events in A-SPACE were controlled by commands sent from the MATLAB audio host via TCP/IP.

2.2 Procedure

Listeners were initially familiarized, outside the ACL, with the sentence lists and visual representations (names and cartoon images) of the six talkers. During subsequent testing, participants were seated at the center of the ACL loudspeaker array. Participants eyes were covered by the HMD, which they wore throughout each run of 60 trials. Ears were uncovered at all times. The HMD presented a uniform light gray background and a visible cursor marking the forward direction.

At the start of each trial, a green square spanning 11° X 11° visual angle appeared at an azimuth and elevation of 0° to indicate the forward-facing “home” position. Participants oriented to and held the home position for 2 seconds to initiate a new trial. A visual marker (white capsule in Fig. 1) then appeared at each of the talker locations selected for that trial. Each marker was randomly assigned a unique azimuth of [±5.625°, ±6.875°, ±28.125°, or ±39.375°] and a random distance of [2.0, 3.3, or 4.6 m]. Markers did not visually overlap. Participants were encouraged to explore the layout, using slow head movements, prior to and during sound presentation, which began 4–7s after the visual markers appeared.

During the training phase of the study, each trial presented one sentence voiced by one of the talkers, along with a visual marker at the same azimuth. Participants were instructed first to localize the direction of the talker (and corresponding marker) by turning the head to position a reticle over the marker and pressing the hand controller’s trigger button. This localization response then triggered the appearance of a virtual button box with six buttons displaying the faces and names assigned to each of the talkers. Participants were next asked to identify the talker by selecting one of the buttons using the controller touchpad. Feedback, in the form of the correct name and face image, was provided after each trial. Participants completed runs of 12 trials each until they reached error-free identification over at least 12 successive trials.

During the main phase, four of the six talkers were randomly selected on each trial. For each selected talker, three sentences were randomly selected, with replacement, from four possibilities. The three sentences were arranged in sequence, separated by intervals of 0.5–1.5s. The resulting stimulus for each talker was presented, with a random initial delay of 0–2s, from a unique location indicated by a visual marker as described above. As indicated by the timeline in Fig. 1, all four talkers were presented concurrently but were not specifically synchronized. Recall that one of the talkers (the target) differed from the others in the room model which generated its reflections. As in the training phase, participants were instructed to locate the target by head pointing and and then to select the virtual button corresponding to the target's identity. The button layout remained constant throughout the experiment to facilitate this task.

3 Results

Localization and talker identification performance were assessed as a function of target room difference in each condition (*Size*+, *Refl*+, and *Refl*-). Psychometric functions appear in Fig. 2, where talker-identification performance is plotted against room parameters: surface reflectivity in conditions *Refl*- and *Refl*+, room dimension for condition *Size*+ (see Table 1). In all conditions, performance ranged from chance (25% correct) when target and standard rooms matched (10m x 10m, $1-\alpha = .5$) to nearly 100% correct for the largest differences.

Thresholds estimated at 75% correct (dashed lines in Fig. 2) are plotted for individual subjects in Fig. 3. Individual thresholds exhibited close correspondence between fits based on localization and talker-identification performance (Fig. 3a), which did not differ significantly ($F_{(1,5)} = 0.008, p = .93$). In Fig. 3b, values are plotted in units of reverberation time RT60, which was closely matched across these conditions (see Table 1). *Size*+ thresholds consistently and significantly exceeded *Refl*+ thresholds ($F_{(1,5)} = 22.48, p < .01$), suggesting that judgments did not simply reflect reverberation time.

4 Discussion

The results of this study demonstrate that listeners can, in fact, discriminate the reverberant characteristics of

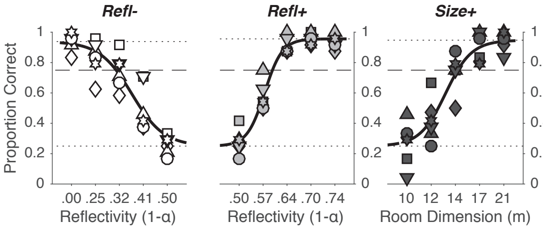

Fig. 2: Psychometric functions plotting percent correct talker identification against room parameters. Symbols plot individual data and curves plot logistic fits [3] to group mean. Dashed lines mark 75% correct level used for threshold calculation.

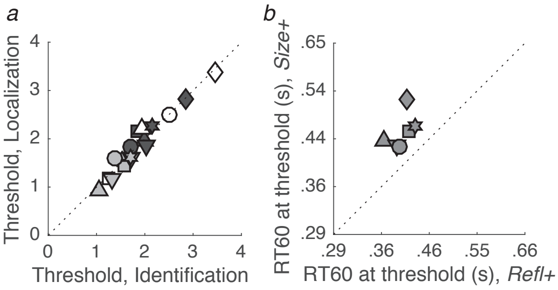

Fig. 3: Threshold estimates from psychometric fits to individual data. a: symbols plot individual thresholds calculated from localization (ordinate) vs. talker-identification performance (abscissa). Values are plotted in units of room index (N in Table 1), where N=0 corresponds to the standard room. Symbols match those of Fig. 2 for participant and condition: *Size*+ (black), *Refl*+ (gray) and *Refl*- (white). b: talker-identification thresholds obtained in condition *Size*+ (vertical) vs. condition *Refl*+.

multiple concurrent talkers in a complex auditory scene when auditory, visual, and dynamic information about talker locations is available. It does not reveal the extent to which participants used these various cues, and in particular does not suggest that sensory immersion is necessary for well-trained listeners to perform the task. Anecdotally, however, participants reported that interacting with the scene helped to focus the spatial impression of each visible talker location. To the extent that visual and interactive aspects of a VR- or AR-based experience help to maintain co-immersion, they should reduce users' sensitivity to target features. A key conclusion of the study is that such sensitivity can be directly quantified in terms of the degree of synthesis discrepancy. In this case, discrimination thresholds were measured for parametric differences in the applied room models; in future cases discrimination could be measured across differences in algorithm type, simulation complexity, etc. An important extension of this work will be to develop versions of this test that can be conducted over headphones using virtual 3D audio and binaural room simulations [4].

The task employed in this study involved two separate judgments on each trial: localization and identification. This approach was adopted in part to mimic natural tasks (i.e. cocktail parties) in which useful information can be obtained by knowing *where* a target sound originates or *what* it designates, and partly to accommodate future testing in populations with reduced spatial and/or linguistic awareness. In the current study, performance on the two tasks was equally good, with few errors on either dimension.

Participants were able to discriminate targets when the room response was enhanced (*Size*+, *Refl*+) or reduced (*Refl*-). Also noted was a consistent difference between thresholds obtained for changes in room size vs surface reflectivity, despite matching the stimuli for reverberation time. This could reflect judgments on a different acoustic basis (e.g., C50), although it is surprising that the ITDG cue—which was not available in *Refl*+ —did not support better performance in the *Size*+ condition. Alternatively, changes in room size may have produced less obvious departures from listeners' expectations, in that everyday listening may feature more variance in room size than in surface characteristics. This result suggests stronger co-immersion when reflection timing and geometry covary with reverberation than when reverberation strength is controlled mainly via amplitude decay.

(Anne)	"Albert threw the mace"
	"Albert threw the mace again"
	"Albert threw the mace up"
	"Albert threw the mason"
(Evan)	"Jill observed the ten then"
	"Jill observed the ten"
	"Jill observed the tendon"
	"Jill observed the ten today"

Table 2: Example sentence lists for two of the talkers.

Finally, in addition to the goals of assessing co-immersion in virtual experiences, the broader impacts of this work suggest uses of immersive simulations to assess spatial hearing, particularly in clinical populations, children, hearding-aid and cochlear-implant users, and aging listeners. The game-like format of VR testing suggests its potential for auditory training, for example to improve spatial awareness in new cochlear implantees or to assess ASA preservation across signal processing approaches, devices, and algorithms.

Acknowledgements

This work was supported by NIH Small Business Technology Transfer (STTR) Grant R41DC16578. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIDCD or the National Institutes of Health. Additional thanks to Julie Stecker for managing the behavioral study and collecting the data.

References

[1] L. S. White, S. King. The EUSTACE speech corpus (http://www.cstr.ed.ac.uk/projects/eustace). Centre for Speech Technology Research, University of Edinburgh, 2003.

[2] G. C. Stecker, T. M. Moore: Reverberation enhances onset dominance in sound localization. J Acoust Soc Am **143** (2018) 786–793.

[3] F. A. Wichmann, N. J. Hill: The psychometric function: I. Fitting, sampling and goodness of fit. Percept Psychophys **63** (2001) 1293–1313.

[4] D. N. Zotkin, R. Duraiswami, L. S. Davis: Rendering localized spatial audio in a virtual auditory space. IEEE Trans Multimedia **6** (2004) 553-564.