Gaze-Contingent Auditory Displays for Improved Spatial Attention in Virtual Reality

MARGARITA VINNIKOV, National Research Council of Canada
ROBERT S. ALLISON and SUZETTE FERNANDES, York University

Virtual reality simulations of group social interactions are important for many applications, including the virtual treatment of social phobias, crowd and group simulation, collaborative virtual environments (VEs), and entertainment. In such scenarios, when compared to the real world, audio cues are often impoverished. As a result, users cannot rely on subtle spatial audio-visual cues that guide attention and enable effective social interactions in real-world situations. We explored whether gaze-contingent audio enhancement techniques driven by inferring audio-visual attention in virtual displays could be used to enable effective communication in cluttered audio VEs. In all of our experiments, we hypothesized that visual attention could be used as a tool to modulate the quality and intensity of sounds from multiple sources to efficiently and naturally select spatial sound sources. For this purpose, we built a gaze-contingent display (GCD) that allowed tracking of a user's gaze in real-time and modifying the volume of the speakers' voices contingent on the current region of overt attention. We compared six different techniques for sound modulation with a base condition providing no attentional modulation of sound. The techniques were compared in terms of source recognition and preference in a set of user studies. Overall, we observed that users liked the ability to control the sounds with their eyes. They felt that a rapid change in attenuation with attention but not the elimination of competing sounds (partial rather than absolute selection) was most natural. In conclusion, audio GCDs offer potential for simulating rich, natural social, and other interactions in VEs. They should be considered for improving both performance and fidelity in applications related to social behaviour scenarios or when the user needs to work with multiple audio sources of information.

Categories and Subject Descriptors: H.5.1 [**Multimedia Information Systems**]: Artificial, augmented, and virtual realities; Audio input/output; Evaluation/methodology; H.5.2 [**User Interfaces**]: Evaluation/methodology; Interaction styles

General Terms: Human Factors, Experimentation, Performance

Additional Key Words and Phrases: Gaze-contingent displays, user experience, visual-audio attention, sound modulation

ACM Reference Format:
Margarita Vinnikov, Robert S. Allison, and Suzette Fernandes. 2017. Gaze-contingent auditory displays for improved spatial attention in virtual reality. ACM Trans. Comput.-Hum. Interact. 24, 3, Article 19 (April 2017), 38 pages.
DOI: http://dx.doi.org/10.1145/3067822

1. INTRODUCTION

When people interact with the real world, they are typically inundated by a cacophony of sounds, sights, smells, and other sensory information. Fortunately, the brain has

R. Alison was supported by the Discovery Grant form NSERC Canada.
Authors' addresses: M. Vinnikov, Flight Research Laboratory, Aerospace Portfolio, National Research Council of Canada, 1200 Montreal Rd., Bldg. U-61, Ottawa ON K1A 0R6; email: margarita.vinnikov@nrc-cnrc.gc.ca; R. S. Allison and S. Fernandes, Centre For Vision Research, York University, 4700 Keele Street, Toronto, Ontario, M3J 1P3; emails: allison@cse.yorku.ca, suzettefernandes@gmail.com.

DOI: http://dx.doi.org/10.1145/3067822

mechanisms that allow people to selectively attend to sources that are the most important in a given situation. This allows people to choose to redirect their attention from one source to another. Compared to these experiences in the real world, our sensory stimulation in virtual reality (VR) is often quite impoverished. A major drawback to multi-modal VR display systems, which stimulate many different senses concurrently, is that they require additional resources, computational power, and consistency between display modes. Because of these challenges, one display mode, normally visual display, is usually well represented, while other modes such as audio and tactile displays are represented with a lower fidelity or not at all. In the current article, we consider whether measures of a user's intent, signalled by their gaze direction, could be used to selectively enable simulation of human spatial auditory capabilities without the requirement for an individualized, high-fidelity spatial audio display.

Specifically, we attempt to simulate the auditory filtering techniques humans naturally demonstrate in the so-called *cocktail party (CP)* effect [Cherry 1953]. The CP effect refers to the ability of a person to attend to a single speaker when exposed to multiple concurrent speakers. The example that inspired the term is a CP, where guests try to participate in conversations in a room full of people speaking. Interestingly, as complex as such a task may be, people still manage to successfully communicate. In the CP effect, listeners demonstrate selective attention, attending to one source of information among many. This ability relies on successful sound discrimination and localization. In the real world, localizing sound depends on binaural [Strutt 1907] and spectral cues [Batteau 1967] unique to each listener. These cues are rarely simulated with high fidelity in computerized audio environments, and even in VR, these are typically highly impoverished compared to real environments and reduced to stereophonic or surround sound presentations. Thus, natural abilities, such as those demonstrated in the CP effect, are significantly limited in VR.

In this article, we will explore cross-modal interaction between visual and audio cues in the common case, where auditory and visual attention are linked (for example, during a conversation at a CP). As a prototypical case, we will examine the utility of gaze-contingent auditory highlighting to simulate the CP effect. The goal is to enable auditory discrimination in cluttered virtual environments (VEs) with performance similar to that found in real-world environments. Specifically, tolerance to audio clutter, as the people's ability to discriminate multiple sound sources drops very quickly with very few sound sources [Miller 1947; Brungart and Simpson 2007]. To motivate and provide the theoretical groundwork for this approach, we will review the relevant literature on attention in auditory processing and review related techniques for auditory highlighting in other fields, such as teleconferencing and desktop human-computer interaction (HCI). We envision that such a system could be beneficial in any VR setup where multiple avatars are involved. For example, this could be a social game or a virtual party or meeting. Specifically, this article will present three experiments, where the first two experiments will compare different techniques to highlight the voice of the speaker who is being looked at by the participant, both objectively in terms of speaker discrimination and subjectively in terms of realism and ease of use. The final experiment will address the question of whether gaze-contingent highlighting has an advantage over a more traditional methods of selection.

2. PREVIOUS WORK

As the focus of this article is auditory attention, it is important to understand what is implied by the term, as well as to discuss the current understanding of human capabilities for auditory attention. The importance of spatial auditory cues for attention has been recognized in other fields, and so we then review previous techniques for auditory highlighting in fields, such as teleconferencing and desktop HCI. The review

then will explain how multi-modal attention can further help with extracting valuable information from the world around a user. Finally, the section will discuss the role of a gaze in selective attention, how gaze-tracking could help in the implementation of a simulated CP effect and what challenges one need to overcome for a successful implementation.

2.1. Auditory and Multi-Modal Selective Attention

The survival of many species directly depends on successful sound discrimination and localization. For humans, each ear receives a sound pressure signal that is the sum of all the signals emanating from different sources and reverberating off surfaces in the environment. Auditory scene analysis relies on the brain exploiting consistencies in the world and between the ears, expectations of source and scene regularities, a grouping of related signals, and other processes and assumptions to segregate sound sources and perceive objects in the world [Shamma et al. 2011]. Successful auditory scene analysis depends on selective attention, although there is debate whether this selective attention acts to select auditory streams or to influence the formation of these streams [Shamma et al. 2011].

In turn, selective attention is affected by (1) different characteristics of the sound environment (open, closed, and furnished); (2) temporal, spatial, spectral, and other similarities and distinctions between competing sources; and (3) unique properties of the sound source (material or animal species) that can help to filter and discriminate spatially distinct audio streams. When dealing with speech, the spectral profile of the speaker's voice (average pitch, speed, sex, and accent of the speaker), prominence of transitions, such as subject matter, voice dynamics, and syntax, can be very important when attempting to attend to different speakers. However, it is not always enough to discriminate between sound signals; it is also important to recognize the location (origin and direction) of sound. Binaural sound localization can be explained by Duplex theory [Strutt 1907] that states that sound source location can be determined from the *interaural time difference (ITD)*, which is the time difference between the sounds reaching each ear and *interaural level differences (ILD)*, which are differences in sound level entering the ears. Such cues are augmented by spectral cues that depend on the shape of the head and outer ear to spectrally shape the sounds arriving at each ear. The combination of these cues can be described by the so-called *head-related transfer function (HRTF)* that describes the transfer function of a sound signal from a spatially localized source as it propagates to the inner ear [Batteau 1967]. Specifically, the HRTF is position, frequency, and distance dependent and describes the interactions of the sound with the listener's pinna (the grooves and notches of the pinna in particular), and to a lesser extent, the head, shoulders/torso.

Even with these capabilities, there are many cases when sounds can be ambiguous. The sound might propagate through objects or reflect from surfaces. One way to disambiguate is by visual observation (multi-modal disambiguation). Hence, visual-audio augmentation plays a very important role in our daily interaction with the world. Numerous studies have shown how visual-audio augmentation can enrich experiences, for instance, when both visual and audio attention are directed at the same spatial source. Other studies have shown how it can sometimes turn into a distraction, for example, when attention is diverted to secondary sound sources, the listener often cannot correctly localize the primary source [Kunka and Kostek 2010; Kunka et al. 2010; Witkin and Leventhal 1952]. Specifically, Maddox et al. [2014] compared the effectiveness of guiding attention with visual and auditory cues and found that visual guidance aided sound discrimination but auditory guidance did not. Such findings provide strong motivation for audio-visual systems that exploit or reproduce these benefits of audio and

multi-modal spatial attention. In the next section, we review prior work towards this goal in related fields.

2.2. Teleoperation, Telepresence, and Video-Conferencing

While the main focus of this article is on immersive VR, other application domains face similar challenges in multi-modal display. Controlling and minimizing resources, especially bandwidth is an important driver in broadcasting, teleoperation, telepresence, and video conferencing [Duchowski 2000; Kortum and Geisler 1996; Stelmach and Tam 1994; Tsumura et al. 1996].

One technique proposed for bandwidth reduction for video streaming in these real-time applications is known as foveation. The idea is to only transmit regions of high interest at the highest resolution, where other areas are presented at a lower resolution (or with more compression). These regions of high interest can be predetermined by complex image analysis algorithms or instead by exploiting eye-tracking technology that can determine regions of interest in real time [Geisler and Perry 1998; Duchowski 2007].

Selection based on user interest can also be used in audio or cross-modal displays. Much of the previous work in this area has been in the context of teleconferencing systems. In the real world, sound segregation and discrimination can be enhanced by localization cues [Cherry 1953]. These natural attentional mechanisms can be used in audio systems to differentiate between several sound sources, to identify various speakers, and to make speech more intelligible [Brungart and Simpson 2005; Drullman and Bronkhorst 2000; Baldis 2001; Goose et al. 2005]. Spatial localization using binaural and spectral audio cues [Billinghurst et al. 1998] allows filtering of irrelevant sounds to focus on just a few competing sound sources [Egan et al. 1954]. For example, Kilgore and Chignel [2003] developed a communications tool called *Vocal Village* to help with voice discrimination by providing spatialized voices in real time. They found that this technique improved the user's ability to discriminate and identify speakers as well as reduced the time required for these tasks. Furthermore, they found that by localizing different sound sources people can process more information when the information is presented from a single audio source.

Some researchers looked at improving telepresence by moving the videoconference into virtual or augmented space [Nakanishi et al. 1996; Benford et al. 1997; Megiddo 2003]. For example, Sellen et al. [1992] developed a system named *Hydra* that positioned all remote participants in virtual monitors arranged around the user [Sellen et al. 1992]. Furthermore, spatial localization can be promoted if users can naturally interact with the 3D space, for example, if they can naturally orient themselves in this space. Billinghurst et al. [1998] provides an example of an augmented reality (AR), where the user can move through the virtual space and attend to multiple conversations with the spatial cues provided through tracking the user's head orientation and location. Billinghurst et al. [1998] built a prototype of such a system, but he did not provide a systematic evaluation of his system. Similarly, Okada et al. [1994] created the *MAJIC* system that simulated a VR space with remote participants projected onto a large-curved screen at natural scale to create an illusion of sitting together around the table. The system supported parallel conversations by allowing an eye-contact between participants. This supported a realistic face-to-face viewing scenario, where the eye-contact cues provided an extra layer of non-verbal communication. In addition, this permitted users to naturally transfer turns and indicated which participants were participating in which conversation. While these two systems captured and displayed real people, Mortlock et al. [1997] used a virtual avatar to represent the people participating in a conference call. Finally, this allowed users to control their view and virtual location and also to establish spatial relations between the speakers.

Deo et al. [2007] compared head-tracking and phone-tracking (i.e, the location and orientation of a mobile device) as a means to spatialize voices of conference participants. They found that both means of tracking allowed for better comprehension of the audio content and better spatial localization of the speakers' locations. As a control condition to determine the effect that volume had on the responses they included a condition where volume was amplified for sources in the direction of the head orientation. While not the main focus of the article, they observed that that simple volume filtering was a very effective approach for improving speech discrimination and spatial localization. This is an important finding, as it suggests that the spatialization techniques can be simplified to volume modulation techniques. However, while correlated with a gaze, head-tracking is imprecise, as it does not take into account the movements of the eyes. For modest changes in gaze, head movements are not required, and thus head-tracking alone can be unnatural as well as imprecise. Furthermore, all or none filtering of sound sources differs greatly from the more subtle attentional selection humans have evolved. Such a system might be beneficial in teleconferencing, where only a single participant should normally be speaking, but the impact of such drastic filtering on preserving a realistic experience in VR is not clear. Hence, we were interested in exploring how the magnitude and time course of attentionally selective volume filtering affects the realism and effectiveness of social communication in VR.

In attentionally selective systems, such as those described in this section, the focus of the user's attention needs to be identified to filter or highlight the information. While this can sometimes be defined by explicit user input or the nature of the task, in many cases, it needs to be inferred from the user's behaviour. In the next section, we review techniques that rely on one of the most powerful and reliable indicators of user attention, their gaze.

2.3. Gaze-Contingent Display (GCD) or Selective Displays

One of the strongest indicators of a person's attentional interest is the gaze direction. Although people can attend to peripheral stimuli (covert attention), typically they look at what interests them (overt attention). Thus, a user's current gaze can be used as a strong indicator of visual attention. Systems that use gaze to indicate attention and modify the display are referred to as Gaze-contingent displays (GCDs). These applications use information acquired from an eye-tracking device and translate this information into a desired output on the display. The term Gaze-contingent display (GCD) is usually used when such transformation happens without a conscious selection of target by the user, while *selective displays* are applications that reflect the intentional selection of a particular item in the virtual or real world (e.g., an 'eye mouse'). Often, the GCD technology extrapolates gaze location in the scene from raw eye position and modifies the corresponding point or object in virtual space accordingly. Such modifications are usually designed to improve visibility and accuracy of the object of interest (assumed to be the object at the point of regard). Although most of the time such displays are based on visual (graphic) processing, they could be extended to modify auditory content as well. An exception is gaze-contingent audio-visual substitution applications that perform sonification to represent data through sound [Twardon et al. 2013; Losing et al. 2014]. Yet, there are almost no GCD systems that utilize multimodal (visual-audio) augmentation rather than the replacement of one sense with the other. One of the very few examples is the system described by Starker and Bolt [1990] that zoomed in and provided detailed audible narration about fixated objects.

A fundamental problem with interfaces based on gaze is the ambiguity of gaze-contingent selection [Hyrskykari et al. 2005], also known as the *Midas touch problem* [Jacob 1991]. In ancient legend, the gift of converting things into gold by touch turns into a curse; similarly, the ability to select things by gaze can be a curse. This can

be attributed to the fact that gaze is not primarily used for explicit pointing. Instead, most of the time, the human eye performs sensory input and eye-movements support this goal with operations, such as scanning and tracking. Furthermore, not all eye movements are voluntary. As a result, gaze-contingent selection can affect areas in a scene that are not regions of attention. Common strategies to avoid the Midas touch problem are the use of intentional blinks, voice [O'Donovan et al. 2009; van der Kamp and Sundstedt 2011], or prolonged dwell time on the area of interest. Unfortunately, these solutions are more applicable when the gaze is used as an explicit selection tool rather than in the context of GCDs. Selection in GCD requires a natural (implicit) approach, so that the user does not need to consciously control their gaze [Tanriverdi and Jacob 2000]. To solve this problem in auditory GCD, Bolt [1981] proposed using attention history statistics to control the volume level in one of the several windows displayed if the user paid extended attention to it for a significant period of time, and to decrease the volume if a given area was only briefly looked at. In other words, gaze history can be summarized with a heat map overlaying the scene as in Kunka and Kostek's [2010] work. This information could then be overlaid with the 3D scene to determine which sound sources should be the most prominent and which should be de-emphasized to the user. In our article, we wanted to explore this even further and see how much of a gaze history needs to be retained for an effective and natural interaction.

Besides indicating attention, gaze and mutual gaze play important roles in conversation and other social interactions. This makes GCD systems particularly attractive for conversational scenarios. According to Hindus et al. [1996], social VR has two major problems that a system developer needs to overcome. First, as in teleconferencing, users face the challenge of discriminating and identifying multiple speakers within the VR space. In the real world, spatialized audio helps to solve this problem, but it is usually not available or impoverished in VR. Second, they need to determine which agents or avatars they can interact with. It is important for the users to know that people are paying attention to them and establish eye-contact [Sellen 1992]. Ho et al. [2015] showed that people utilize very distinct gaze patterns to indicate the beginning and end of a speaking turn. For instance, they observed that speakers look directly at the listener once they are ready to allow the listener to speak. Gaze behaviour is thus an important aspect of conversation and in the present study, we implicitly rely on this natural relationship between gaze and conversational intent. Others have explicitly used mutual gaze for augmentation. For example, Yonezawa et al. [2007] developed a stuffed robotic teddy bear that responded when a user looked at it. They observed that eye-contact between the robot and the user made the user feel more favourably toward the teddy bear compared to situations when there was no eye-contact. The user liked the robot even more if the eye-contact was combined with joint attention, where the robot looked at the same direction as the user. Similarly, Vidal et al. [2015] found that when avatars responded to eye-contact, the users experienced increased feelings of immersion and awareness of their own eye-movements. Castellina and Corno [2008] looked at gaze as part of multi-modal interactions in 3D VR. Their results suggested that gaze-controlled interactions provide a better VR experience. Consistent with this conclusion, Bente et al. [2007] reported that longer gaze interactions significantly improved participants' perception of co-presence. To deal with the loss of mutual gaze information in a teleconferencing scenario, Vertegaal [1999] used an eye-tracker to convey regions of attention (focus of attention and workspace activities) of the user to other participants. By incorporating gaze, Vertegaal was able to convey the relationship between speakers (knowing who is talking to whom, and who is talking about what). Hence, we think it is important to explore how people feel when the system responds to their non-verbal cues, such as fixating different avatars.

3. GENERAL METHODOLOGY AND EXPERIMENTAL APPROACH

As mentioned earlier, in a normal 3D world, the user has rich spatial cues to help segregate and attend to objects and auditory streams. These are usually absent in a VE aside from basic (and usually uncalibrated) binaural cues from stereophonic audio. Consequently, in a series of experiments described in this article, we were interested in exploring the relationship between visual attention (as measured through gaze patterns) and the aural information acquisition of speech. Specifically, the intent was to identify the primary design choices that can lead to better audio-visual human-computer interfaces. More specifically, can GCD be used to improve the way people acquire auditory information in simulations of real-life scenarios? Is it sufficient to manipulate the volume and quality of audio input contingent on the user's gaze-location to improve audio voice discrimination in 3D scenes? Similar to the graphical studies of Kistler et al. [2010] and Pelechano et al. [2008], our system simulated a VE with multiple avatars located at different positions in the virtual space. We were interested in an environment that would closely resemble a *CP* scenario, in other words, where multiple avatars speak at the same time. To resemble real scenarios, the intensity of the sound projected from each avatar in the scene depended on two factors: the avatar's spatial location and the area of interest, which was driven by the user's gaze. We used the user's gaze to enhance the audio streams in a task-dependent manner to ease speech recognition and source identification in a cluttered audio scene. We hypothesized that gaze-contingent attentional selection could augment or substitute for normal spatial attention underlying the real-life CP phenomenon and related effects. We used gaze as an indicator of current conversational interest and highlighted the attended source while de-emphasizing distracting speakers through volume manipulation. This gaze-contingent enhancement allowed simulation of the attentional effects of spatial auditory attention without full high-fidelity spatial sound simulation. We assessed these techniques both in terms of the impact on speech recognition and the naturalness of the virtual environment (VE) interface. Two types of experimental design were used. First, we were interested in acquiring an overall user impression of our techniques along different aspects associated with VE. Second, we were interested in directly comparing the techniques to assess the relative effectiveness and naturalness of the interfaces. Finally, we want to verify that GCD can be used as preferred method of interaction with VR agents in social group interactions.

3.1. System Setup

Unity3D was used to create and present the VE for the experiment. Besides the rapid 3D scene development, scripting options, and experimental flexibility provided by this platform, Unity3D has support for manipulating sounds, modelling sound propagation in the 3D environment and customized sound manipulation. Specifically, it incorporates a virtual microphone (*AudioListener*) that represents the listener. This object acquires all the sounds propagated from different sound sources within the virtual scene with regards to the spatial location of each sound source relative to its (AudioListener) location. The AudioListener was located at the main camera (i.e., the simulated user location). Unity also correctly distributed sounds to generate proper audio for the stereo headphones.

As was noted by Witkin and Leventhal [1952], sound proximity is biased by the presence of faces. Choosing the right facial representation (avatar) is very important. To create animated avatars, we used Crazytalk7 Professional version. This application generated animations for characters, including their lip and facial movements, based on the audio and text files corresponding to their assigned speech segments. The fact that the user's view was fixed in the virtual scene during the experiment allowed us

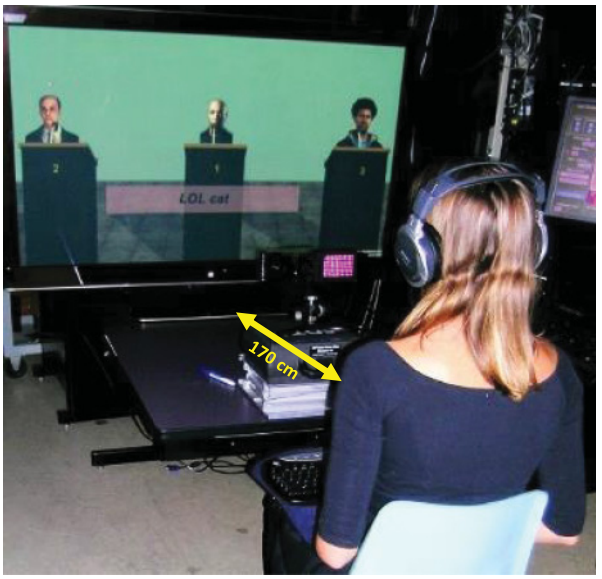

Fig. 1. Experimental setup. The user sat in front of a TV screen while her eyes were tracked by EyeLink 1000.

to use avatars with 2.5D representations (with no back meshes), which significantly reduced rendering complexity.

3.2. Apparatus

Visual displays were generated on a desktop computer with AMD FirePro W9000 FireGL, Windows 7 Enterprise, Intel® Core™ CPU, and 3.50GHz, 3.50GB Ram. The stimulus was presented on a Panasonic Viera TCP54VT25 54inch diagonal, HDTV plasma display with a pixel resolution of 1920H $\times$ 1080V and a refresh rate of 60Hz. During the experiment, the screen was viewed binocularly at a distance of 1.6–1.7m from the user. This viewing arrangement provided a window into the VE with a horizontal visual angle of $39.5^\circ \pm 1.5^\circ$ (as shown in Figure 1). A real-time gaze-contingent system was assembled by incorporating an EyeLink 1000 (SR Research) eye-tracker, which supports online tracking of the user's eye and head position. The sampling rate was 500Hz. To provide maximum comfort to the user and range of head motion, we used a desktop mount that allowed free head movements over a volume of 7920cm^3. As a result, the user's head location was tracked by localizing an identifiable marker placed on their forehead in the eye-tracker video stream. The tracker has an average head and eye (gaze) tracking accuracy of 0.5° and a precision of 1.0° in the remote mode we used. We measured the end-to-end latency at 18.33 ± 5.5ms using the technique described in Vinnikov and Allison [2013]. The experiment was conducted in a darkened room to maximize gaze-tracking performance. To achieve realistic 3D sound, we used Sony MDR-XD200 Stereo headphones that transmitted stereo sound to the user. Finally, for the purpose of the third experiment we also used a Targus® Wireless Laser Presenter with Cursor Control.

An EyeLink 1000 built-in calibration procedure was always performed before each experimental block. The calibration involved sequentially fixating nine points displayed on the screen in a pseudo-random order. Each calibration procedure was followed by the EyeLink 1000 built-in validation procedure, where the participant successively fixated a series of additional nine targets. The error was calculated as the difference

between target and estimated eye position. At the end of the validation procedure, an accuracy parameter was displayed to indicate whether the calibration was successful or not. In the case that the average tracking error was larger than 1.0° in the display area, then the calibration was attempted again. If the required calibration accuracy could not be obtained, the participant was dismissed with credit for participation in the experiment.

4. EXPERIMENT 1

The first experiment evaluated different variations of gaze-contingent auditory highlighting in the context of the CP effect and subjective user ratings. Overall, the hypothesis was that such gaze-contingent attentional selection could augment or substitute for normal spatial attention underlying the CP phenomenon and related effects. We assessed how the parameters of the GCD audio enhancement affected the ability to recognize speech in a cluttered audio environment. Specifically, we compared six techniques that differed in the amount and time course of the attenuation of the background or the amplification of the attended speaker. In addition to examining different techniques, we varied the number of speakers. We asked users to subjectively rate each technique on the task and on measures related to social presence and immersion. Furthermore, we had performance-based dependent variables, specifically, the accuracy of identifying the correct speaker and the response time between start of a trial and identification of the speaker. Finally, we recorded both the raw eye movements and the objects of interest for each given instance.

4.1. Methods

4.1.1. Task. The users were asked to imagine that they were participating in a discussion group, where all of their group members were virtual characters. Each trial consisted of a short discussion session in a VE. During the discussion session, each avatar spoke about a randomly selected topic. Each speech segment was a repeated monologue that was approximately one-minute long. At the beginning of each trial (discussion session), the user was given a keyword, which was the topic of one of the speeches presented by the avatars. The users' task was to identify which avatar was speaking about the indicated topic. The users were instructed to identify the correct avatar in the most accurate and timely manner. As such, each of these discussion sessions is referred to as a *speaker identification trial*.

4.1.2. Stimuli. Six different gaze-contingent auditory techniques were implemented (Table I). The first technique was an *absolute sound cut-off* (A), where the participant was able to hear only one speaker at a time corresponding to their current conversational interest, while all other sound sources were muted. The second condition was a *partial sound cut-off* (P). This technique presented the fixated sound source at its peak volume, while all other sound sources were reduced to 20% of nominal volume (programmable volume supported by Unity[1] for all participants.). The remaining techniques had filtered sound modulation, based on the time a particular target was fixated. In other words, the longer a user looked at a target the louder it got, while other targets became quieter. Consequently, we had *fast sound modulation up and down* (FF) and *slow sound modulation up and down* (SS). Finally, we had volume that increased quickly with attention but faded slowly when unattended (FS) as well as the converse of slow increase with rapid fade (SF). Fast modulation from maximum volume to minimum volume was 1.13 seconds, while slow modulation was 9 seconds. For the four cases (FF, SS, SF, and FS), the background drop-off was capped to 20% of maximum

[1]The volume of the headphones on the volume control options was setup to 50%.

Table I. Six Different Gaze-contingent Auditory Techniques

Technique name	Foreground	Background	Transition between speakers
Absolute sound cut-off (A)	Speaker of interest at 100% volume.	All other speakers are muted.	Instantaneous.
Partial sound cut-off (P)	Speaker of interest at 100% volume.	All other speakers are at 50% volume.	Instantaneous.
Fast modulation for foreground speaker and fast volume drop for background (FF)	Speaker of interest reaches 100% volume over time.	All other speakers drop volume to 50% volume over time.	Fast (maximum duration of 1.3 seconds) volume amplification for the speaker of interests and slow (maximum duration of 9 seconds) volume drop for unattended speakers.
Slow modulation for foreground speaker and slow volume drop for background (SS)	Speaker of interest reaches 100% volume over time.	All other speakers drop volume to 50% volume over time.	Slow volume amplification for the speaker of interests and slow volume drop for unattended speakers.
Fast modulation for foreground speaker and slow volume drop for background (FS)	Speaker of interest reaches 100% volume over time.	All other speakers drop volume to 50% volume over time.	Fast volume amplification for the speaker of interests and slow volume drop for unattended speakers.
Slow modulation for foreground speaker and fast volume drop for background (SF)	Speaker of interest reaches 100% volume over time.	All other speakers drop volume to 50% volume over time.	Slow volume amplification for the speaker of interests and fast volume drop for unattended speakers.

volume (programmable volume), just as in condition P, and the foreground volume had the potential to reach 100% of maximum volume (programmable volume). All trials started with 50% of maximum volume (programmable volume). The modulation was exponential ($1 - e^{\frac{-t}{\tau}}$) to model the perceptual sound attenuation that would occur in the real world [Stevens 1955].

Seven distinct avatars were used in our experiment (see Figure 2). Despite their distinct facial features, care was taken to ensure that each avatar's face subtended the same visual angle (5.55°). The avatars spoke behind a wooden stand that was marked with the avatar's ID number. We expanded the bounding ellipsoid to include both the avatar's upper body and the wooden stand, because there are instances when people pay attention to what the speaker is saying but no longer look at his/her face [Võ et al. 2012]. The number of avatars varied: Each audio technique was tested with three, five, and seven avatars. To control for the size of grouping (the user's required field of view), care was taken to ensure that the first and the last speakers always appeared at the same locations on the screen; the rest of the avatars were evenly distributed out in the remaining space. Thus, we varied both the number and density of speakers in the scene.

The audio content was also carefully selected. While only male voices were used (it has been shown that people can easily discriminate between male and female voices), there was a variety of speakers to simulate a realistic scenario. We also used a wide range of topics, where each speaker spoke about a distinct topic (a given term or

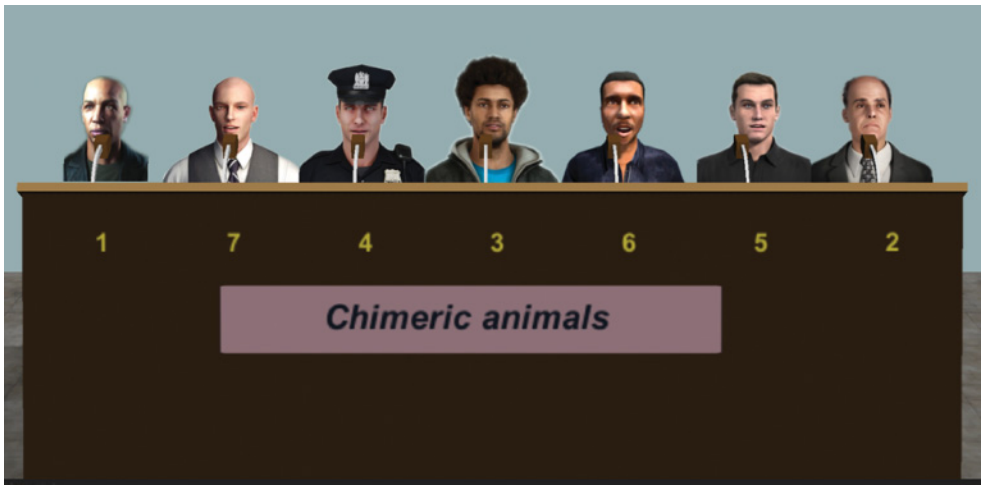

Fig. 2. Example of a speaker identification trial with seven speakers.

situation). These clips were obtained from various pod-casting websites[2]). The keyword and keyword synonyms were repeated multiple times throughout each monologue. An example of text for the term 'allowance' is provided below:

> Many children first learn the value of money by receiving an *allowance*. The purpose is to let children learn from experience at an age when financial mistakes are not very costly. The amount of money that parents give to their children to spend as they wish differs from family to family. Timing is another consideration. Some children get a weekly *allowance*. Others get a monthly *allowance*. In any case, parents should make clear what, if anything, the child is expected to pay for with the money. At first, young children may spend all of their *allowance* soon after they receive it. If they do this, they will learn the hard way that spending must be done within a budget. Parents are usually advised not to offer more money until the next *allowance*.

Finally, we made sure that there were no similar topics (or similar vocabulary) used simultaneously on any given trial.

The system augmented the basic Unity3D spatial sound modelling that affect the sound volume for each speaker. The basic Unity3D speech sound model is based on physics and the location of the virtual object relative to the user. This information determined the sound that the user would hear. For example, objects presented farther away resulted in quieter sounds to the listener's ear as opposed to objects that were closer, assuming both objects had equivalent intensity at their source. This is conventional sound source modelling for 3D VEs and was of limited use to the stationary user in this experiment with approximately equidistant speakers (the speakers were standing at the same virtual distance away from the participant along the z-axis but at different distances along the x-axis). On top of this conventional virtual sound environment, we introduced a second sound model that enhanced or substituted the

[2](a) ESL Podcast – http://www.ivoox.com/podcast-esl-podcast_sq_f1350_1.html; (b) Learn Out Loud – http://www.learnoutloud.com/Podcast-Directory/Travel; (c) Encyclopedia of Life – http://podcast.eol.org/podcast; (d) English Listening Lesson Library Online – http://www.elllo.org/english/0701.htm; (e) VOA Learning English – http://learningenglish.voanews.com/; (f) Culips English Learning Podcast – http://culips.com; (g) The English Desk – http://englishdesk.blogspot.ca/; (h) ESL Business News – http://www.eslbusinessnews.com/.

user's auditory spatial attention and thus required determining the focus of the user's attention. For instance, avatars that the user looked at projected louder messages than those that were not of interest to the user. This was achieved by real-time gaze-tracking [Marmitt and Duchowski 2002].

4.1.3. Procedure. Each experiment started with users completing in a short demographics questionnaire, followed by an online audio test that checked their ability to hear over a range of frequencies (500–4,000Hz) [AuD OnLine Hearing Test]. The users then performed a short English comprehension test to identify baseline for overall language proficiency and to train users for the main experiment. Users that had significant difficulty understanding spoken English were excluded. After completing the test, users were shown a demo trial and were given instruction about how to respond during the trial.

The main experiment followed and was divided into six experimental blocks. Each block consisted of a calibration procedure, six speaker identification trials and a post-block questionnaire. During each trial, the user listened to a group of avatars all simultaneously speaking on independent topics, as described in Section 4.1.1. Once a user was able to identify the correct avatar, he or she had to press the keyboard space bar. This action prompted to a new screen, where the user was required to type in the ID number of the avatar that was identified as the target speaker. Users could change their answers in case of any typing error until they registered it by pressing the space bar once more. After each block of six speaker identification trials, users completed a questionnaire that assessed their user experience during the block. This sequence continued for the duration of the experiment until all six blocks were completed. At the end of the experiment, users were debriefed and all of their questions were answered and any additional comments about the experiment were recorded.

Each of the six blocks was associated with a different one of the audio augmentation techniques discussed in Section 4.1.2. The administration order of the blocks was randomized based on a Latin Square design. Each block consisted of six trials. The trials differed in the number of avatars speaking with two repeat trials for each avatar set size in each block. Trial order was randomized for each user within each block. Similarly keywords, sounds and speakers' locations were individually randomized. Overall, each subject participated in 36 trials over the duration of the experiment (six audio conditions by three avatar set size conditions, each repeated twice).

4.1.4. Participants. We recruited 38 users for this study; however, 10 were dismissed due to poor performance on the English language proficiency screening and poor calibration. Therefore, 28 users took part in the full experiment (14 females and 14 males, ranging in age from 18 to 38, average age 25.25). All users had uncorrected distance visual acuity of 20/30 or better, with good hearing in both ears. They were all highly proficient in the English language. Written informed consent was obtained from all users in accordance with a protocol approved by the York University Ethics Board.

4.1.5. Questionnaire Design. We developed our post-block questionnaire based on the Biocca et al. [2003] social presence questionnaire. All questions were administrated with an online questionnaire and used a 7-point Likert scale to rate each question, with 1 as *never* and 7 as *always*. The questions are presented in Table II and are grouped into several categories. The first category was the *Spatial Perception* category and included Questions 1 and 2. We were interested in these questions because, in the real world, to change the volume of the sound, one has to change the distance from the sound source. In other words, one has to approach the speaker or move away from the speaker to hear him better or to stop listening. Consequently, we expected that in their replies users would indicate whether or not they needed the sound volume to

Table II. List of Questions in the Post-block Questionnaire (The List was Shuffled Before it Given to Users)

Spatial perception	Q1.	I wanted to listen to speakers from a closer distance.
	Q2.	I wanted to make specific sounds louder or softer.
VR immersion	Q3.	I found it easy to forget that I am listening to virtual speakers rather than real people.
	Q4.	I found it easy to forget that I was watching a display.
	Q5.	I felt as if the speakers and I were in different places rather than the same room.
	Q6.	To what extent did you feel consciously aware of being in the real world?
Visual perception	Q7.	I was distracted by the visual quality.
	Q8.	I liked the visual quality of the speakers.
	Q9.	I wanted to see the speakers more closely.
Audio attention	Q10.	I was easily distracted by other speakers. When listening to a given speaker.
	Q11.	Speakers were able to communicate their stories clearly to me.
	Q12.	I could pay close attention to a speaker I was interested in.
	Q13.	I was able to pay attention to multiple speakers simultaneously.
Action–Reaction	Q14.	I felt I knew what was going to happen next (in terms of sound).
	Q15.	Speakers paid close attention to me.
	Q16.	The speakers were affected by who I paid attention to.
Overall experience	Q17.	It was easy to adapt to the current audio technique.
	Q18.	The task I had to perform was difficult.
	Q19.	I enjoyed listening to the current block.
	Q20.	The block was natural.

change; we expected that this might be the case for the techniques with slow volume modulation (SF, SS, and FS). In addition, we wanted to see if the people immersed themselves in the VE and had the urge to physically influence the sound quality. We asked more direct questions on this topic in the second category related to *VR Immersion* composed of Questions 3, 4, 5, and 6. These questions addressed the user's awareness of the discrepancy of the fusion between the real world and the virtual scene. We expected that the unnatural/uncomfortable trials would make users more aware of the real world. The third category was *Visual Perception* and included Questions 7, 8, and 9. These questions looked at how, if at all, the auditory techniques influenced the users' perception of visual experiences such as image quality. We hypothesized that there should be an effect of the techniques as the auditory quality was guided by the user's visual attention. The fourth category was *Auditory Attention,* and it included Questions 10, 11, 12, and 13, which inquired about whether the techniques made it easier or more difficult for the user to direct her/his auditory attention to specific speakers. The fifth category, *Action–Reaction*, included Questions 14, 15, and 16 that dealt with the contingency aspect of our system, and checked whether people could predict, manipulate, and get immersed in the virtual experience. We expected that a well-calibrated and responsive system would create a positive and interactive experience. Finally, the last section *Overall Experience* included Questions 17, 18, 19,

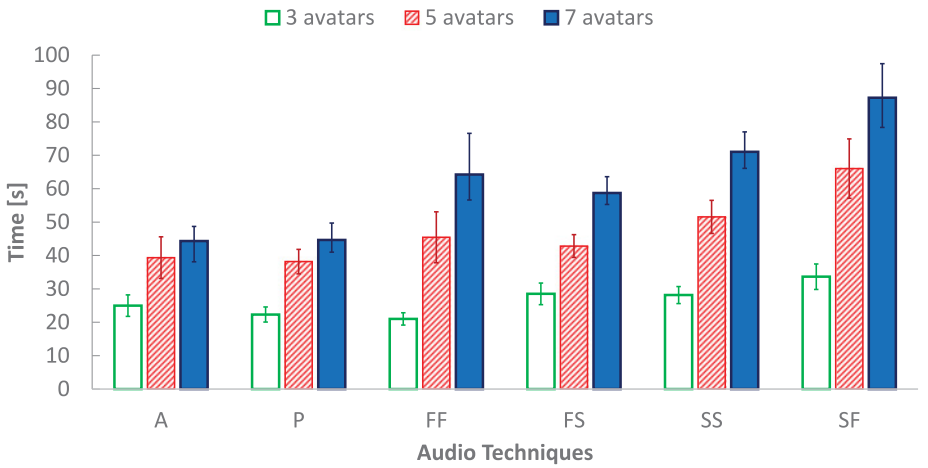

Fig. 3. Mean keyword detection time as a function of technique and number of speakers, averaged across participants. The error bars represent the standard error of the mean for each condition. Absolute unattended sound cut-off (A); partial unattended sound cut-off (P); fast volume amplification for speaker of interest and fast unattended volume drop (FF); slow volume amplification for speaker of interest and slow unattended volume drop (SS); fast volume amplification for speaker of interest and slow unattended volume drop (FS); slow volume amplification for speaker of interest and fast unattended volume drop (SF).

and 20. These questions asked about the general user experience, and we expected to see major differences between the implemented techniques on these questions.

4.2. Results

We present the results for each dependent variable separately below.

4.2.1. Psychophysical Results.

Errors. Users were required to identify the speaker corresponding to a given conversational topic but in some cases could not do so or identified the incorrect avatar. The average number of these errors was 5.88 (out of 36 trials) over the entire experiment averaged across users. Error rate was similar across techniques and only the SF condition had a slightly higher error rate (12.35% versus 8.13% across the other five conditions). As it was not clear what would be the user's performance without augmented audio, it was planned to include a non-gaze-contingent condition; however, after a pilot study, this condition was not possible for our users to do in a reasonable amount of time, if at all. Without the ability to utilize attention to listen to a particular speaker more closely, users performed very poorly with only 42.0% correct identifications on average versus the 91.05% correct (across all six conditions) that was achieved in the present study.

Detection Time. The keyword detection (target speaker identification) time across techniques and number of speakers is shown in Figure 3. Analysis with a two-way ANOVA revealed significant main effects for the number of avatars ($F(1.84, 101.13) = 55.21$, $p < 0.001$, $\eta^2 = 0.501$) and audio technique ($F(2.971, 173.78) = 10.80$, $p < 0.001$, $\eta^2 = 0.164$) on the task performance time. Results also revealed a significant interaction between the number of speakers and audio technique ($F(2.175, 387.58) = 2.175$, $p < 0.05$, $\eta^2 = 0.038$). Detection time was longer for a larger group of avatars than for a smaller group, but the effects of audio technique were larger with an increased number of avatars speaking. Pairwise comparisons (with Bonferroni correction for multiple comparisons) revealed significant differences between all combinations of number of speakers (3, 5, and 7: all $p < 0.001$). Similarly, pairwise comparisons revealed

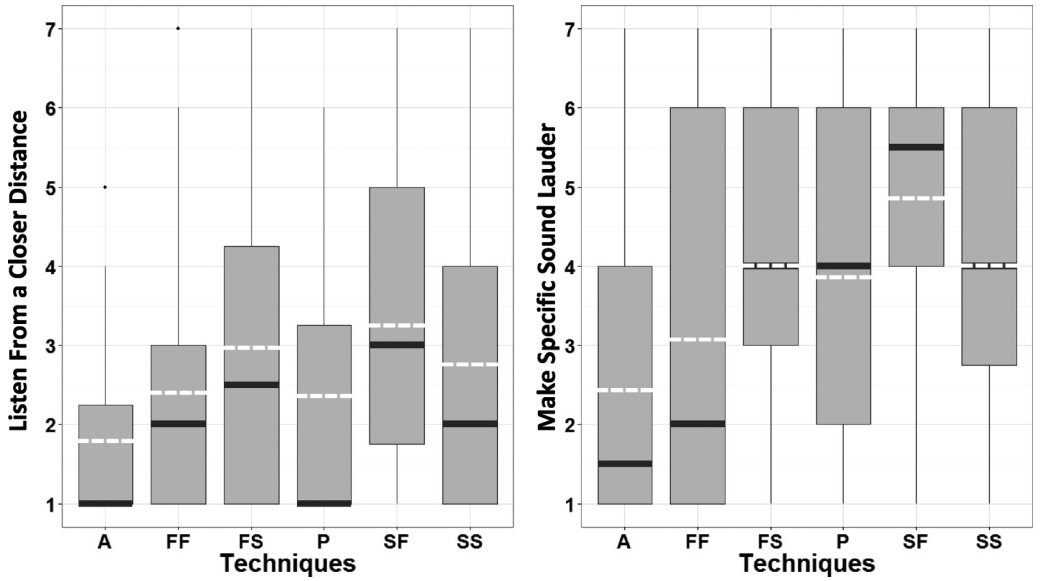

Fig. 4. Likert results for spatial perception questions (with 1 meaning never and 7 always): Question 1 – I wanted to listen to speakers from a closer distance (Left); Question 2 – I wanted to make specific sounds louder or softer (Right). Each box represents the lower and upper quartiles (top side –25% and bottom side – 75%), the dark horizontal line inside the box indicates the median, the white dashed line the mean, the extent of the whiskers show the minimum and maximum values that are not outliers, and the stand alone points are outliers (more or less than 1.5 times of upper or lower quartile, respectively.

significant differences between A and SS techniques ($p < 0.05$), between the A and SF ($p < 0.001$), between the P and SS ($p < 0.001$), and between the P and SF ($p < 0.001$). Detection times were longest in the SS and SF conditions.

4.2.2. Questionnaire Data. We will present the data from the post-block questionnaires in terms of groupings of related questions, as discussed in Section 4.1.5. Effects of augmentation technique were assessed with Friedman tests, and post-hoc pairwise comparisons between conditions were performed using Wilcoxon signed-rank tests with Bonferroni correction.

Spatial Perception. The distribution of questionnaire responses on Question 1 and Question 2 are presented in Figure 4. The desire to approach the speakers (Question 1) or to adjust the volume of the sources (Question 2) varied with augmentation technique ($\chi^2(5) = 18.22, p < 0.003$ and $\chi^2(5) = 36.68, p < 0.001$, respectively). Participants expressed the strongest desire to approach the speakers in the SF condition but the difference was only significantly different from the A condition ($W = 1, Z = -3.26, p < 0.05, r = 0.44$) and no other pairwise differences were significant. Users wanted to change the volume of sounds significantly less in the A condition than in the other conditions except for FF (n.s.) and significantly less in the FF condition than in SF ($p = 0.001$). Inspection of the figure shows that conditions with a slow increase of the attended to speaker (SF and SS) seemed more troublesome in terms of volume than the equivalent fast increasing conditions (FF and FS). A Tukey post-hoc analysis with Bonferroni adjustment confirmed that the conditions with fast increase in foreground volume were significantly different than those with slow increase ($p = 0.002$).

VR Immersion. There were no significant differences between techniques for the questions related to immersion (Questions 3–6). The average score for Question 3,

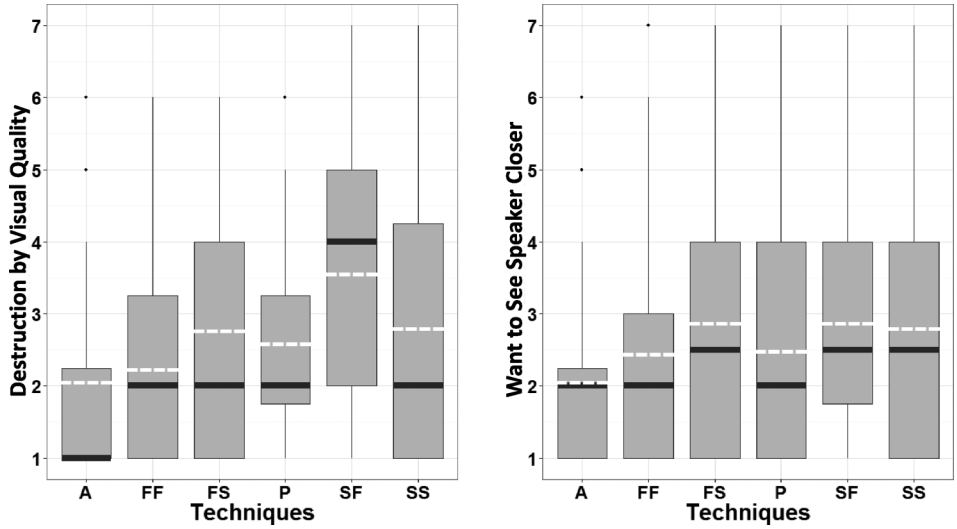

Fig. 5. Likert results for visual perception questions (with 1 meaning never and 7 always): Question 7 – I was distracted by the visual quality (Left); Question 9 – I wanted to see the speakers more closely (Right).

4, and 5 were 3.4, 3.2, and 3.2, respectively, which indicates a *neutral* attitude. For Question 6, the mean response was 5.0, which suggests that people were aware of the real world around them. Immersion seemed to be weak, presumably reflecting the fish-tank VR setup (as opposed to a more immersive setup).

Visual Perception. The average score on Question 8 ('I liked the visual quality of the speakers') was 4.54, which indicated the users' overall satisfaction with the visual content but there were no significant differences between conditions. In contrast, there was a significant effect of condition for Question 7 and 9 (Figure 5) ($\chi^2(5) = 20.61$, $p < 0.001$ and $\chi^2(5) = 12.56$, $p < 0.05$, respectively). Post-hoc analysis indicated that visual distraction was less in technique A compared to SF ($p = 0.04, r = 0.41$).

Auditory Attention. The distribution of questionnaire responses related to audio attention are presented in Figure 6. There was a significant effect of technique for Questions 10, 11, and 12 ($\chi^2(5) = 27.45, p < 0.001$; $\chi^2(5) = 18.21, p < 0.01$; and $\chi^2(5) = 18.37, p < 0.01$, respectively). Unattended speakers were less distracting (Question 10) in the A condition than in all other conditions except FF. For Question 11 and Question 12 responses indicated that participants had more trouble communicating with and attending to speakers in the SF condition (significantly lower average Likert responses in than in the A, FS, and P conditions for both questions as well as FF for question 11). In addition, participants reported better ability to attend to the speaker when volume was increased rapidly rather than slowly for the speaker of interest (FF and FS versus SF and SS $p = 0.019$).

Action–Reaction. There were significant effects of technique for Questions 14 ('I felt I knew what was going to happen next (in terms of sound)'), 15 ('Speakers paid close attention to me'), and 16 ('The speakers were affected by who I paid attention to') ($\chi^2(5) = 12.12$, $p < 0.05$; $\chi^2(5) = 19.76$, $p < 0.01$; $\chi^2(5) = 12.50$, $p < 0.05$, respectively) as shown in Figure 7. Once again, condition SF was least preferred and significantly different than condition A ($p = 0.05$ and $p = 0.04$, for Question 14 and 15, respectively). For Question 15, participants also responded that speakers seemed to pay less attention to them in condition SF than condition FF ($p = 0.05$) and in conditions with slow (SS

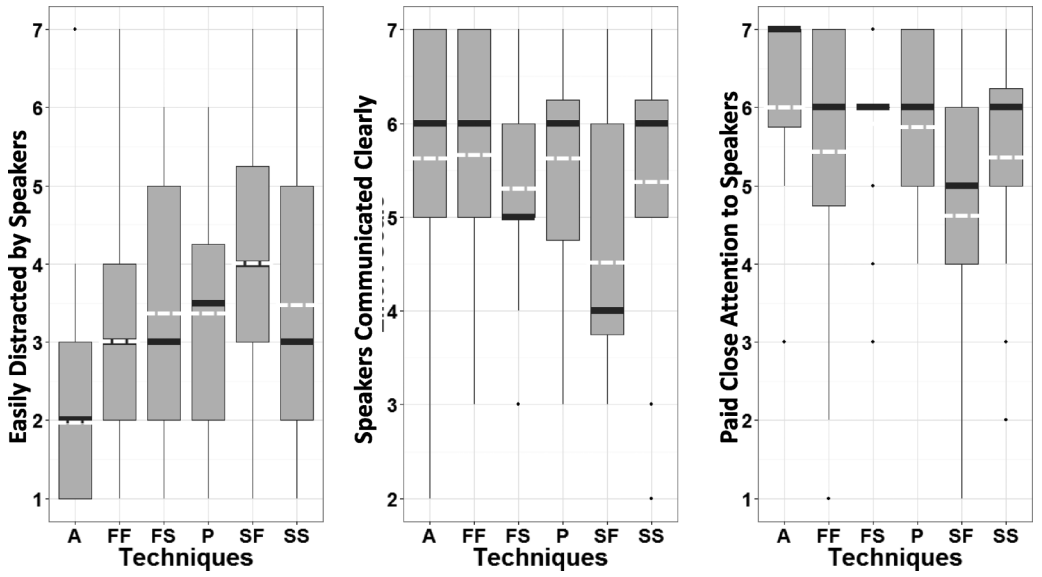

Fig. 6. Likert results for audio attention questions (with 1 meaning never and 7 always): Question 10 – I was easily distracted by other speakers, when listening to a given speaker (Left); Question 11 – Speakers were able to communicate their stories clearly to me (Middle); Question 12 – I could pay close attention to a speaker I was interested in (Right).

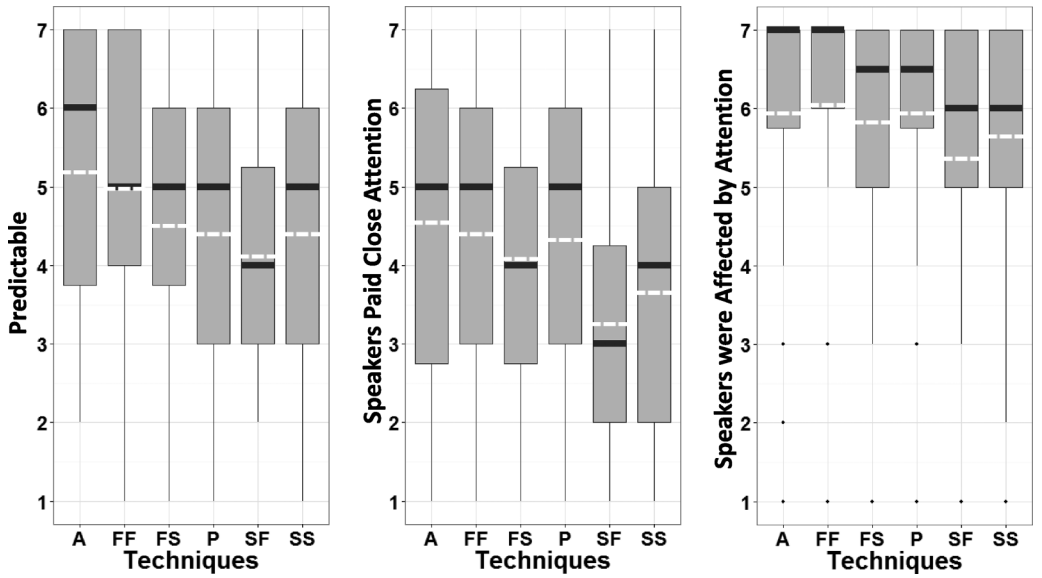

Fig. 7. The distribution of questionnaire responses for 'action–reaction' questions (with 1 meaning never and 7 always): Question 14 – I felt I knew what was going to happen next (Left); Question 15 – Speakers paid close attention to me (Middle); Question 16 – The speakers were affected by who I paid attention to (Right).

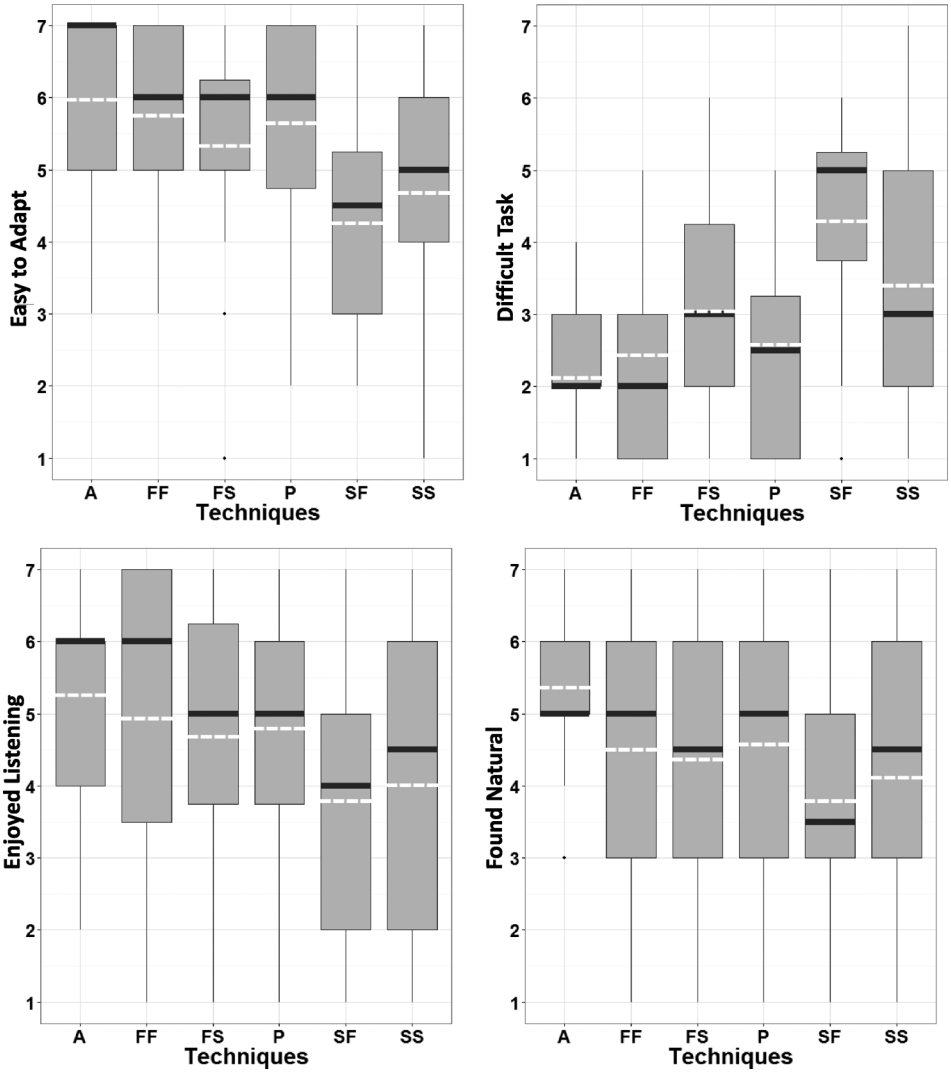

Fig. 8. Likert results for overall experience questions (with 1 meaning never and 7 always): Question 17 – It was easy to adapt to the current audio technique (Top left); Question 18 – The task I had to perform was difficult (Top right); Question 19 – I enjoyed listening to the current block (Bottom left); Question 20 – The block was natural (Bottom right).

and SF) compared to fast (FF and FS) volume increase for the speaker of interest ($p = 0.005$).

Overall Experience. There were significant effects of technique for all four questions related to overall experience (Figure 8) (Q6: $\chi^2(5) = 34.77$, $p < 0.001$; Q18: $\chi^2(5) = 40.82$, $p < 0.001$; Q19: $\chi^2(5) = 20.0138$, $p < 0.01$; Q20: $\chi^2(5) = 22.29$, $p < 0.001$).

Generally, there were lower ratings for conditions with slow volume increase in the foreground (SF and SS) compared to fast or instant switch to the voice of the speaker of interest (A, FF, FS, and P). For all four questions, Tukey post-hoc analysis with Bonferroni adjustment showed that people favoured the condition without background (A) relative to the rest of the conditions considered as a group ($p < 0.001$, except for

Table III. A Two-way Repeated Measures ANOVA for Effect of Number of Avatars and User's Visual Behaviour (Mean Fixation Duration, Fixation Redirection, and Longest Fixation Time) on the UA, SA, and CA

	Average fixation duration			Fixation redirection to a new speaker			Longest fixation time		
	$F(2, 110)$	p	η_p^2	$F(2, 275)$	p	η_p^2	$F(2, 110)$	p	η_p^2
UA	30.65	<0.001	0.36	55.58	<0.001	0.50	–	–	–
SA	12.38	<0.001	0.18	6.20	0.004	0.10	3.41	0.04	0.06
CA	12.25	<0.001	0.18	5.28	0.009	0.09	1.69	0.19	0.03

Table IV. A Two-way Repeated Measures ANOVA for Effect of Technique and User's Visual Behaviour (Mean Fixation Duration, Fixation Redirection, and Longest Fixation Time) on the SA and CA

	Total time fixating			Longest fixation time		
	$F(2, 275)$	p	η_p^2	$F(2, 275)$	p	η_p^2
SA	5.17	<0.001	0.09	11.07	<0.001	0.17
CA	5.30	<0.001	0.09	10.28	<0.001	0.16

Question 18, which was marginal with $p = 0.07$). Rapid volume increase of the speaker (FS and FF or P and A) was generally preferred to slow volume increase (SS and SF); these differences were significant for Questions 17 ($p < 0.001$), 18 ($p < 0.001$, note this question is negatively worded), 19 ($p = 0.001$), and marginally significant for Question 20 ($p = 0.07$).

4.2.3. Gaze Analysis. To analyse the gaze data, we looked at the eye positions sampled during the experiment. In addition to the raw gaze positions relative to the display, we also recorded all the occasions that the user's gaze intercepted with a speaker or the pulpit in front of the speaker. This information was readily available as it was used to determine what sound sources to modulate. As a result, we could calculate the total time the user fixated a specific speaker or when there was a fixation redirection from one speaker to another.

For the gaze analysis, we identified three categories of speakers: unclassified speakers (UA) that could be any speaker presented during a trial, speakers that the user selected as his/her answer (SA) and speakers that were actually the correct speakers (CA). In many cases, the last two categories would refer to the same speaker, but not always. We then looked at different behaviours that the users exhibited relative to these three types of speakers. We ran a two-way repeated measures ANOVA on the eye-data and looked at two factors—audio technique and number of avatars—and their interaction. In Tables III and IV, we report the significant results for all factors (we did not report interactions as they were not significant).

Average fixation duration is presented in Figure 9. From the figure, and confirmed by the statistical analysis, one can see that the average fixation duration was highly dependent on the number of avatars but not the GCD technique. Post-hoc comparisons primarily indicated differences between conditions with three avatars compared to five ($p < 0.05$ for SA and CA; $p < 0.001$ for UA) or seven avatars ($p < 0.001$ for SA, CA, and UA). In the general case (UA), there was also a significant difference between five and seven simultaneous speakers ($p < 0.001$).

On the other hand, the total time a user spent fixating the selected or correct avatar (sum of all fixation durations on the target over the trial), depended more on the technique type than the number of concurrent speakers (Figure 10) with more time spent fixating the selected/correct speaker in the SF and SS conditions. Post-hoc comparisons indicated significant differences between SS and A (SA: $p = 0.04$; CA: $p = 0.05$), P (SA: $p = 0.02$; CA: $p = 0.003$) and FF (SA and CA: $p = 0.02$) conditions with no evidence of any difference between the SS and SF conditions ($p = 1.0$ for both SA and CA).

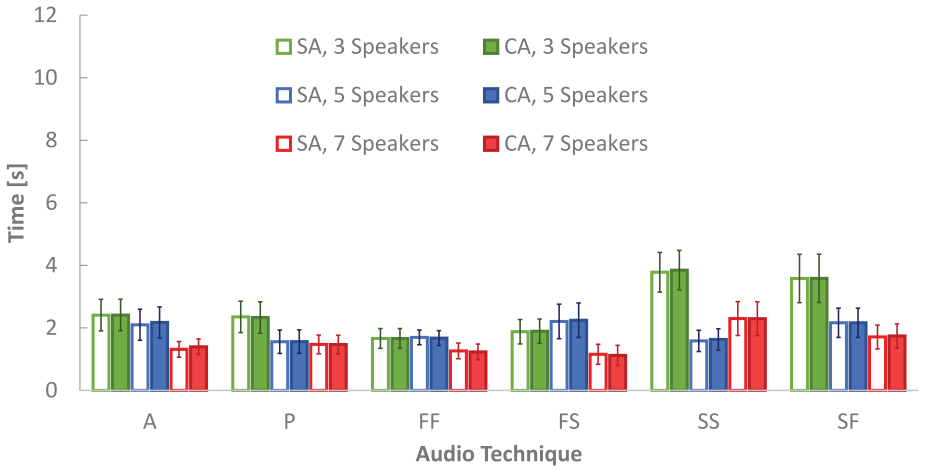

Fig. 9. Mean fixation duration spent fixating the selected and correct speaker as a function of technique and number of concurrent speakers, averaged across users. The error bars represent the standard error of the mean for each condition.

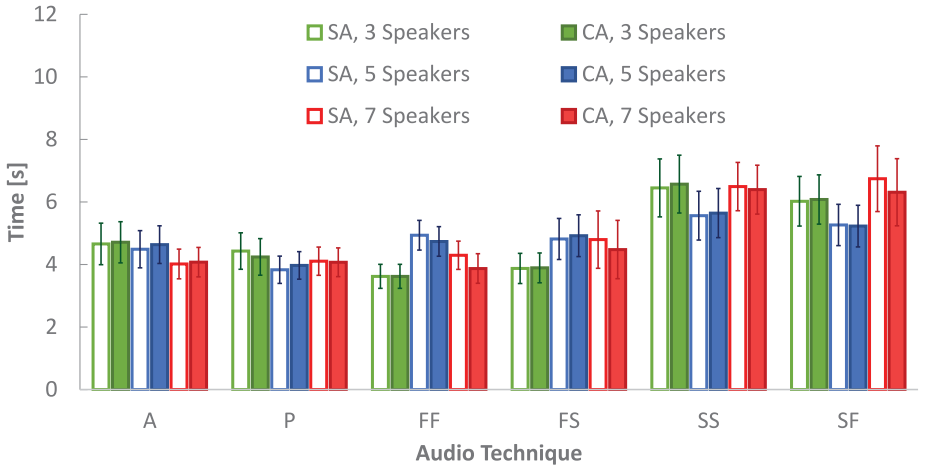

Fig. 10. Mean total time (in a trial) that the user spent looking at the selected speaker as a function of technique and number of concurrent speakers, averaged across users. The error bars represent the standard error of the mean for each condition.

The mean number of fixations that were redirected to a given speaker type in a trial is presented in Figure 11. From the statistical analysis, one can conclude that the number of fixations redirected to another speaker was a function of number of avatars. The number of fixation redirections increased with number of avatars in general ($p < 0.001$ for all conditions of 3, 5, and 7). On the other hand, users performed significantly fewer redirection from the speaker when there were three speakers versus seven speakers for selected or correct avatar ($p < 0.001$).

Finally, the longest fixation duration spent on the selected or correct speaker was influenced by the audio technique used. Specifically, for the longest fixation duration spent on the correct or selected speaker, post-hoc comparisons indicated significant differences between SF and SS compared to the rest of the conditions ($p < 0.01$). Yet, there were no significant differences between SS and SF. Although there was a

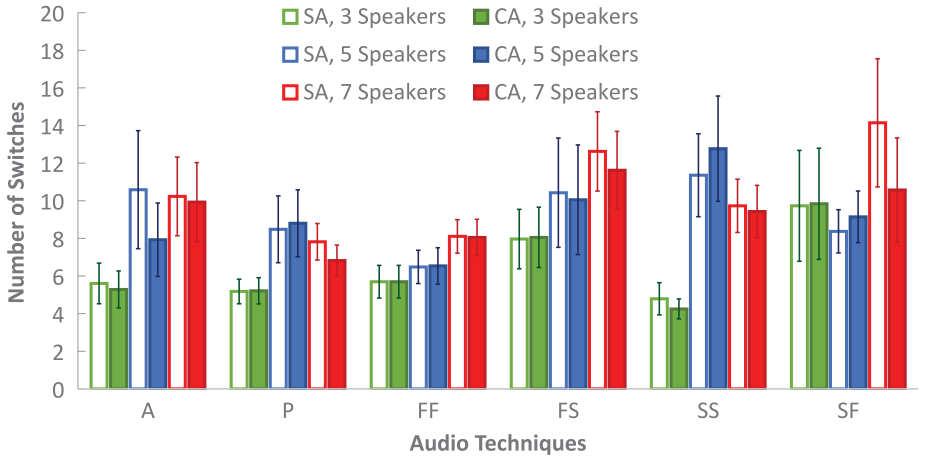

Fig. 11. Mean number of times that fixation was redirected from speaker to another in a trial as a function of technique and number of concurrent speakers, averaged across users. The error bars represent the standard error of the mean for each condition.

significant effect of number of speakers for longest fixation duration spent on selected speaker (Table III), post-hoc analysis did not reveal any specific significant differences between conditions.

4.3. Discussion

4.3.1. Psychophysical Results. The number of errors was on average 8.84% and users reported that the task was relatively simple. As predicted, the time to complete each trial was proportional to the number of avatars (and hence the set of alternatives over which the user had to search). This is congruent to the *CP literature* [Miller 1947; Brungart and Simpson 2007], which has identified seven speakers as a limiting number of concurrent speakers for differentiating the voices. For example, in his study, Miller [1947] showed that eight different voices produced very a strong masking effect. Due to the fact that sounds were modulated in our trial, i.e., not uniform, and that there were never more than seven concurrent speakers, this threshold was likely not reached. In contrast, as noted above the task was very difficult to perform without any modulation or augmentation.

We also observed that the time to identify the correct speaker varied with the technique used during the block. Trials for the condition, where the unattended speakers were rapidly attenuated but the speaker of interest was slowly enhanced, took the longest to complete. In contrast, trials where the voice of the main speaker was amplified rapidly or instantaneously, took the shortest time to complete. Thus, it appears that performance was most enhanced by accentuating the attended speaker than by other manipulations.

4.3.2. Questionnaire Data.

Perception. The general subjective success of our system was demonstrated by user responses to Question 2 ('I wanted to make specific sounds louder or softer'). As expected, since we adjusted the volume of attended and unattended speakers differently for different conditions, it was natural that the users would want to make some sounds louder or softer. Specifically, the significant difference between conditions where the attended speaker was rapidly amplified compared to where it was slowly modulated, highlighted the perceptual impact of the techniques. What was interesting to observe

was that there were significant differences between responses for Question 2, which asked about the user's desire to come closer to or move farther away from the speaker. In other words, did users want to change their spatial distance relative to the sound sources (speakers)? As expected, the most significant difference was between the SF and A conditions, which are the extremes in terms of noise level (none versus an indistinguishable background noise immediately after a change in attention). Participants, thus, felt little need to change position in the A condition but wished to be able to do so, on average, in the SF condition.

VR Immersion and Perception. There was no significant difference between the conditions on questions related to VR immersion, as people felt neutral about the virtual speakers, watching a display or virtual space in all cases. Although the display was large and filled a significant field of view, it is perhaps not surprising that sense of immersion was modest, since nowadays people are very accustomed to animated characters on a big TV screen. Interestingly, in Question 11, people were more aware of being in real world than immersed in the simulated VE. This might be attributed the fact that they were forced to sit in relatively fixed position, and they had to answer the question after, rather than during, the block (i.e., after returning to the 'real world').

When asked about the visual quality of the speakers (Questions 1, 7, and 8), users provided ratings that indicated that they were pleased with overall quality of the experimental scene and there was no significant difference in image quality across the conditions (there were no objective differences in image quality of course). Yet, when users were asked about their satisfaction with visual quality in general, their satisfaction differed significantly between condition A and FS, indicating that they did associate the quality of sound with visual quality. This phenomenon has been previously reported [Rimell and Owen 2000; Frater et al. 2001; Winkler and Faller 2005; Mastoropoulou et al. 2005; Rimell et al. 2008; Cullen et al. 2012], specifically it has been established that quality perception in one modality, for example audio, can impact the perception of another modality, for example visual quality. Finally, when asked, if they wanted to see the avatars more closely, users had significant differences in their replies for condition A and SF, indicating again that they associated visual and auditory quality and also supporting our hypotheses that people's natural spatial association of distance with volume would promote a desire to move closer when speech recognition was difficult.

Audio Attention and Action–Reaction. As discussed, it was very important for us to see users' ratings of techniques to the audio-attention and action–reaction, as these categories touch upon factors crucial in evaluating GCD applications. In particular, it was important to ensure that the behavioural consequences with GCD were as predicted, in other words, did not cause any side effects due to latency or other limitations [Parkhurst and Niebur 2002]. Specifically, we observed that instantaneous or fast modulation of sound, as expected, supported the users' ability to pay close attention to speakers of interest, while delayed modulation led to the opposite effect. Similarly, people found that background speakers were distracting in the SF condition and they had a harder time acquiring clear stories with the techniques that had slow modulation of the foreground sound. These observations emphasize the importance of building GCD systems with minimal, if any, latency [Vinnikov and Allison 2013]. Similarly, the different techniques controlling the sounds influenced how well users could predict what will happen next. For example, in conditions, where the foreground sound changed instantaneously or rapidly, users felt better in control and were immersed more in the environment. Nevertheless, overall, the responses to the questions on audio attention and action–reaction confirm that our techniques indeed produced the desired CP effect even in the SF and SS conditions.

Overall Experience. As expected, users found the SS and SF techniques to be the most challenging to adapt to, the most difficult to operate and the least enjoyable. Nevertheless, the responses were leaning more towards positive than negative rating. Finally, when asked how natural the techniques were, all techniques were rated as natural. The most highly rated technique was the A technique, which is not in reality most natural as in the real world people cannot mute background speakers. Once again, SF was rated as the most unnatural technique. It is possible that people operationalize natural in terms of how easy it is to interact with each technique. We believe that this requires further investigation. A more detailed study regarding what people think is easy and natural warranted (Section 5).

4.3.3. Gaze-Data. In terms of gaze data, an interesting pattern emerged. Averaged fixation duration on the selected and correct speaker in each trial was the highest when there were only three concurrent speakers; this was particularly evident in the SS and SF cases. This could be explained by the fact that users had to divide their attention less between speakers when there were fewer speakers. In term of the total time spent on each speaker, users fixated selected or correct speakers the longest in the SS and SF cases. This is because the user had to wait longer to hear the speaker of interest at the maximum volume. This was not reflected in the average fixation time per speaker, but in the longest fixation they made to the correct or selected speakers. It is possible to assume that the average time was low because of the large number of attention switches or that some fixations were very short. Yet, the total number of attentional switches depended on the number of speakers in the scene rather than technique used. The search space depended on the number of speakers, and thus, users made one and two fixations per avatar on average. It is possible that if a more complex sound modulation was implemented, the user would have been able to listen to several speakers at the same time, hence reducing the number of switches. In addition, there was a difference in the number of fixations redirected to the correct versus the selected speaker. Specifically, the number of redirected fixations was higher for the selected speaker than the correct speaker for the five or seven concurrent speakers. It is likely, that when unsure, users selected the speaker that they attended to the most.

5. EXPERIMENT 2

Following the first experiment, we were interested in comparing and assessing the realism and ease of use of the proposed techniques (Questions 17, 18, and 20) more directly. Specifically, we had users explicitly compare techniques, while discriminating between five simultaneous speakers, as we have established that this number of speakers is already a challenging condition. We added a base case of no gaze-contingent augmentation, where we did not modulate the sound level of attended or background speakers. To compare the effectiveness of the techniques more precisely, we controlled the duration of each trial. This allowed us to avoid ceiling level performance on the speech discrimination task, and thus, we could use error rate as a measure of task performance. We hypothesized that when directly asked to compare techniques, people would be able to accurately identify which techniques are closer to what they will hear in the real world and disassociate the complexity of the test they had to perform from the realism question. Specifically, we expected that users will find the non-gaze-contingent technique the most realistic and the technique without speakers in the background as the most unreal technique. We also predicted that they would find conditions with fast attenuation of the background and rapid enhancement of the speaker as the easiest to use, and this would be reflected in reduced error rate.

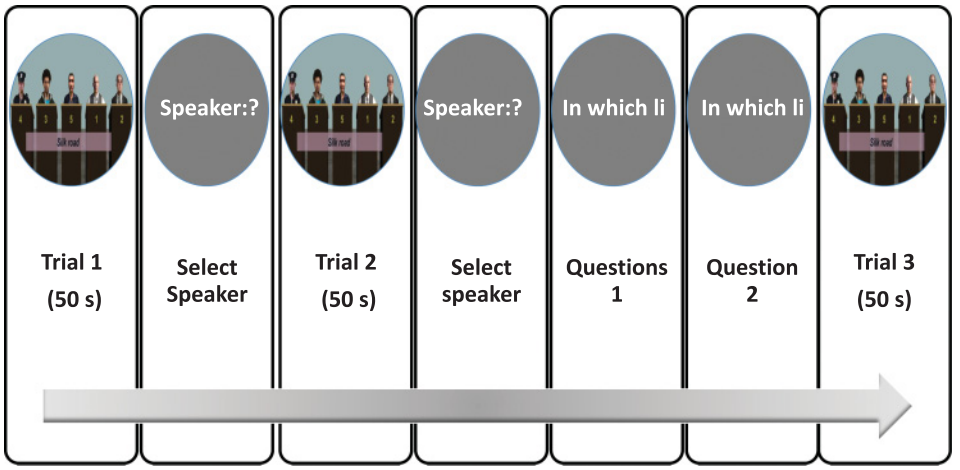

Fig. 12. Procedure for Experiment 2.

5.1. Methods

5.1.1. Task. The speech discrimination task from Experiment 1 was used. As before, users had to listen to a short discussion session and identify the speaker who was discussing the identified topic. The only difference was that, while trial duration was unlimited in the first experiment (trials ended when the user was ready to respond), trials in the current experiment always ended after 50 seconds. In addition, users were asked to compare consecutively presented conditions. As a result, after every two trials, users answered two questions:

—In which listening condition did the conversation feel more natural/realistic?
—In which listening condition was it easiest to attend to a speaker?

Figure 12 shows the sequence for each pair of trials. This sequence was followed through the entire experiment. To ensure properly maintained calibration, the experiment was divided into four blocks of six comparison trials (with two discrimination tasks per trial). Therefore, each user was exposed to 24 trials (48 speech discriminations) in total. Assignment of trials to blocks was randomized for each user, as was the order of the conditions.

5.1.2. Procedure. As in Experiment 1, each session started with users completing in a short demographics questionnaire, which was followed by an online audio test and a short English comprehension tests. After completing the tests, users were shown a demo trial and were given instruction about how to respond during the trial. This was followed by the experimental blocks. At the end of the experiment, users were debriefed and all of their questions were answered as well as any additional comments about the experiment recorded.

5.1.3. Stimuli. In this experiment, we used the same set of animated avatars and spatial layout as in Experiment 1. However, this time, there were always five concurrent speakers. We also used the same corpus of audio clips as well, because a different set of users was recruited for this study. Two groups of users were used to compare different sets of techniques from the previous study. In addition, we also added a non-gaze-contingent condition (N):

(1) Group 1: users had to compare N, A, P, and SF
(2) Group 2: users had to compare FS, SS, FF, and SF

For the first the group, we selected four techniques to compare, each of which varied in the degree of relative enhancement of the speaker and showed significant differences between each other for Experiment 1. Specifically, the two techniques that had most different responses were A and SF. We decided to include N as it was the nominal base case for a typical VE, and we observed that it was the most difficult case during a pilot study. Finally, P was conceptually a different case from A as it maintained speakers at the background while A did not. On the other hand, in Experiment 1, we observed that there was a significant difference between techniques based on sound modulation in the background and foreground and between the different speeds. Therefore, we used FS, SS, FF, and SF trials in the second group to compare these dynamics. Finally, the repeatability of the experiment could be verified, since FF was very close to P condition and FS was used in both groups. Using two groups of participants allowed for the experiment to be run in a single session of modest length and ensured that no participant would hear any speech sample more than once in the experiment.

5.1.4. Participants. We recruited two groups of users from an Undergraduate Research Participant Pool. The first group included 19 users (18 females and 1 male, ranging in age from 17 to 32, average age 18.89). Their English proficiency scores ranged from 86% to 100.00%, with a mean a score of 98%. The second group included 21 users (14 females and 7 males, ranging in age from 17 to 22, average age 18.94). Their English proficiency scores were ranged from 86% to 100.00%, with mean score of 98%. All users had uncorrected distance visual acuity of 20/30 or better, with good hearing in both ears. Written informed consent was obtained from all users in accordance with a protocol approved by the York University Ethics Board.

5.2. Results

5.2.1. Error Rate. Unlike the first experiment, where the unlimited task execution time resulted in very few errors, this experiment had a limited trial duration, and hence, we expected to see a significant difference in error rate between conditions. Indeed, with one-way repeated-measures ANOVA, we found a significant main effect of technique used for both group 1 ($F(3, 54) = 45.18, p < 0.001, \eta_p^2 = 0.72$) and 2 ($F(3, 60) = 9.52, p < 0.001, \eta_p^2 = 0.32$). The distribution of percentage correct responses under different conditions is shown in Figure 13. Overall, one can see that users had the most accurate performance with A, P, FF, and FS techniques and the worst performance with N. Post-hoc paired comparisons using a Wilcoxon signed rank test with a Bonferroni correction were used to confirm differences between the techniques.

In group 1, the A and P techniques had the lowest error rates and were not significantly different from each other ($p = 1.0$). The error rate in the N condition was very high and significantly worse than the other three conditions (all $p < 0.01$). The SF condition had an intermediate error rate that was significantly higher than the A and P conditions ($p = 0.04$ and $p = 0.006$, respectively). Thus, all the augmentation techniques improved recognition performance compared to no augmentation but the condition with gradual attended speaker enhancement was more error prone than the instantaneous A and P conditions.

Conditions for group 2 were arranged to look at the effects of temporal dynamics on the augmentation. While the SF condition showed enhancement, whereby users were not as error prone as N (from group 1), it resulted in significantly more errors than the FS ($p = 0.04$, compared to SF) or FF conditions ($p = 0.008$ and $p = 0.01$, compared to SS and SF, respectively). Combined with the results for group 1, which showed low error in the instantaneous A and P cases, this suggests that the most important aspect of the augmentation is rapid enhancement of the attended speaker. The speed of attenuation of the background seemed to be of lesser importance, since there were no

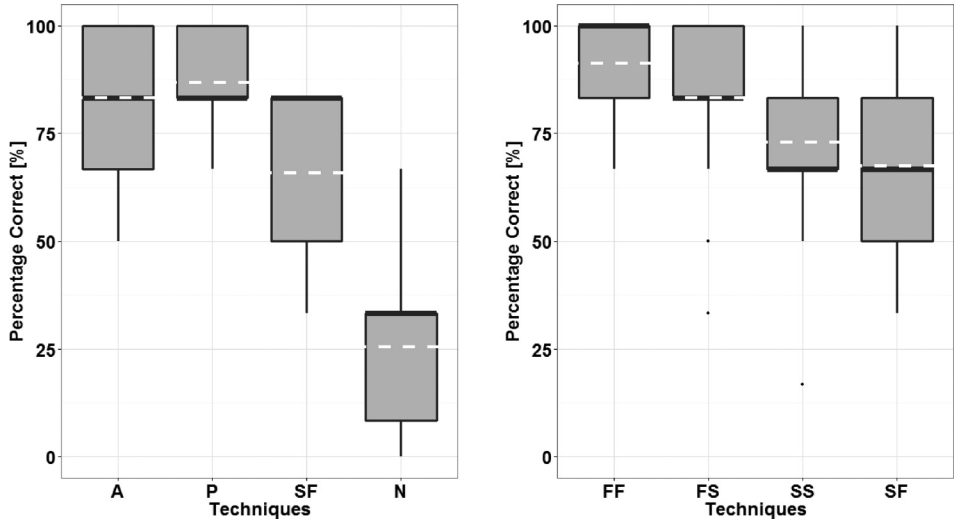

Fig. 13. Percentage correct for speaker identification under different GCD techniques for two test groups: Group 1 (Left) and Group 2 (Right).

Table V. The Chart Shows Estimated Weights of Audio Conditions in Determining Preferences for Q1 (Realism) for Both Groups

Group 1				
	N	SF	P	A
N	0	0.33**	0.35 **	0.30*
SF	−0.33**	0	0.03	0.03
P	−0.35**	−0.03	0	−0.05
A	−0.30*	0.03	0.05	0

Group 2				
	SF	SS	FS	FF
SF	0	0.02	0	0.12
SS	−0.02	0	−0.02	0.10
FS	0	0.02	0	0.12
FF	−0.12	−0.10	−0.12	0

Each row shows the contribution of each audio technique (Columns) to preference relative to the base case (1st Column). A positive estimated weight implies that the case is more likely to be preferred. Significance of each contribution is shown by stars ($p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$).

significant differences between the FS and FF conditions ($p = 0.51$) or between the SF and SS conditions ($p = 1.00$). The statistical analyses therefore confirm that the most important factor in determining the number of errors was the speed of modulation of the speaker of interest.

5.2.2. Preferences. To better understand the relationship between conditions, we performed a paired comparison analysis of the choice data. We used a Turner and Firth Model [Turner and Firth 2007, 2012] that fits generalized non-linear models (gnm) using an over-parameterized representation.

The weight estimates for both groups are presented in Tables V (Q1) and VI (Q2). The tables show the estimated weights from the fitted Poisson (log) model with the analysis repeated four times with each condition treated, in turn, as the base case (in separate rows). The estimated weights show the contribution of the augmentation technique to preference relative to the base case; a positive estimated weight implies that the case is more likely to be preferred in comparisons than the base case.

Figure 14 shows preference responses for combination of conditions, polled over all users. The y-axis represents the percentage of times the choice listed first in the pair of conditions was selected (each pair was presented twice during the experiment, but

Table VI. The Chart Shows Estimated Weights of Audio Conditions in Determining Preferences for Q2 (Easier) for Both Groups

Group 1				
	N	SF	P	A
N	0	0.64***	0.96***	1.47***
SF	−0.64***	0	0.32*	0.83***
P	−0.96***	−0.32*	0	0.51***
A	−1.47***	−0.83***	−0.51***	0

Group 2				
	SF	SS	FS	FF
SF	0	0.07	0.38**	0.58***
SS	−0.07	0	0.32**	0.51***
FS	−0.38**	−0.32**	0	0.20.
FF	−0.58***	−0.51***	−0.20.	0

Each row shows the contribution of each audio technique (Columns) to preference relative to the base case (1st Column). A positive estimated weight implies that the case is more likely to be preferred. Significance of each contribution is shown by stars ($p < 0.1$, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$).

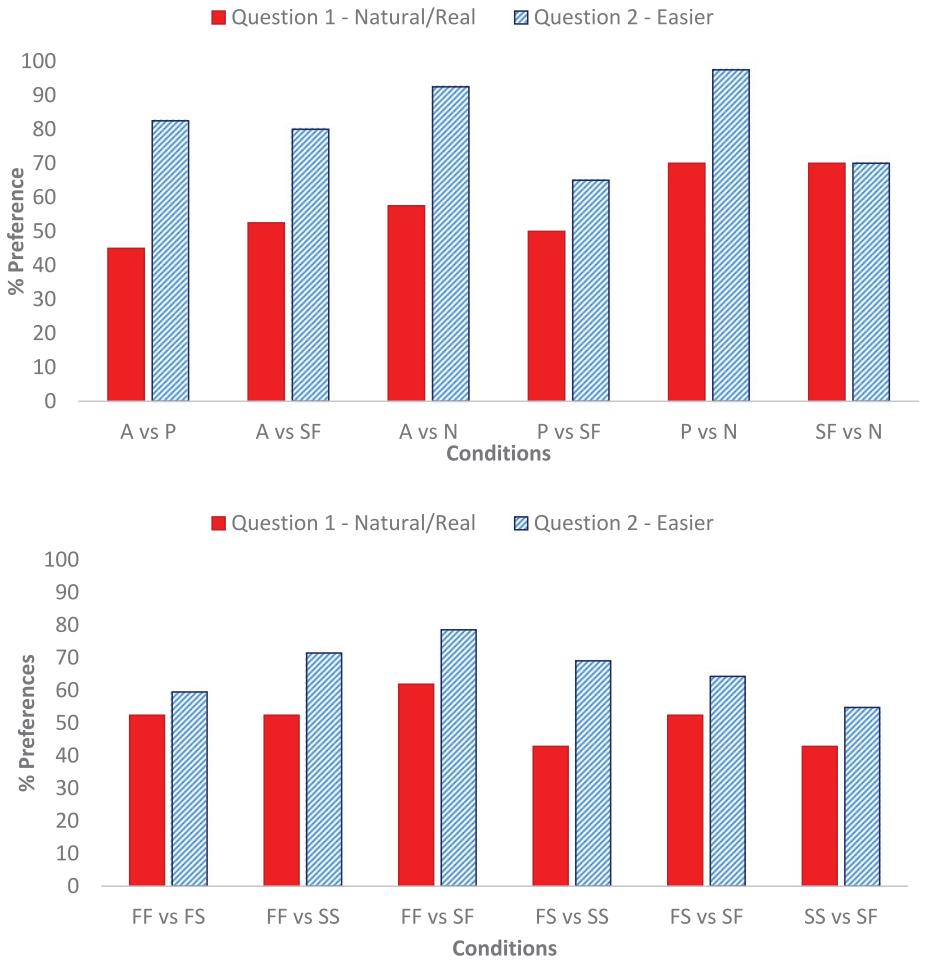

Fig. 14. The mean preference pooled over participants for group 1 (top) and group 2 (bottom). Preferences are coded as the percentage of times the first condition in the pair listed was preferred over the second (note that first and second do not correspond to presentation order, which was randomized).

in reversed order). The score of 50% indicated no preference between the conditions on average.

From the figure (Figure 14) and the table of weights (Table V), one can see that, overall, there were not very strong preferences between conditions in term of realism. The only significant difference was in group 1, between the augmented audio techniques and the base case (N). However, contrary to expectations, people thought the augmented audio techniques were more realistic than the base case. The weight estimates and the preference ratings for the techniques with partially muted speakers (P and SF) were larger (more realistic) than for the A condition but this difference was not significant. There were no significant differences in realism between the different temporal dynamic conditions compared in group 2.

On the other hand, users had strong preferences between audio techniques when asked which of the conditions made the task easier. Group 1 compared various levels of augmentation and the A condition was judged easiest followed by the P, then SF and then the N condition. Thus, the task was judged significantly easier as the degree of enhancement of the attended speaker increased (from none, N, to absolute, A). This is reflected in significant differences in weights between all of these conditions. In terms of the temporal dynamics, the conditions with the rapid enhancement of the attended speaker (FS and FF) were reported to be easier than those with slow enhancement (SS and SF). For conditions differing only in the rate of attenuation of the background, when the enhancement of the attended stimulus was slow, there were no significant differences (SF versus SS), and when the attended stimulus was rapidly enhanced, there was only a marginally significant preference for a faster offset (FF versus FS). Thus, perceived task performance seemed to be mainly affected by the rate of enhancement of the attended speaker and the effect of background modulation on perceived task performance was small.

5.3. Discussion

The results showed that accuracy in identifying the correct speaker was highly related to the rate of sound modulation of the speaker of interest. This was evident in both the error measures and the subjective responses that indicated better performance for conditions with rapid or instantaneous enhancement of the currently fixated speaker. Perceived and objective task performance was primarily affected by the rate of enhancement of the attended speaker and the effect of background modulation on perceived task performance was small.

Rapid emphasis of the attended speaker is presumably beneficial, since it allows for the immediate analysis of the selected audio stream with higher signal-to-noise ratio. Therefore, a performance improvement in these conditions was to be expected as the user could pick up the keywords that would have helped them to identify the speaker earlier. We were concerned that a rapid onset would either be (a) distracting if one made a chance glance to an avatar they were not actually interested in listening to (e.g., the Midas touch problem) or (b) be perceived as unnatural and distracting. Although a slower volume adjustment was arguably a more natural condition, it was not perceived as such and did not produce more accurate identification.

The synchronous phase of modulation of background and foreground sounds also played an important role, especially in the slow modulation case. We expected that users would have the hardest time with the SS case, where there was slow increase in foreground speaker volume and slow volume decrease for previously heard speakers. However, this technique was not as detrimental as the SF condition, where the foreground volume was slowly increased and the background was dropped down rather quickly. We suspect that this was a more difficult case as it did not provide the user the

ability to relate previously attended sounds relative to the newly enhanced voice. As a result, the user has to start the sound discrimination task from scratch.

Contrary to expectations, people thought that the augmented audio techniques were more realistic than the base case (N). The N case is the most realistic in the sense that the audio streams were unmodified. However, the scenario is unrealistic in that the richer spatial audio cues the user could normally rely on in the real world were not present and there was no evidence of a CP effect (participants could not attend well to a single conversation in this condition). Thus, it may be that user felt the augmented cases were more realistic since they provided the ability to attend to a single speaker in a CP scenario that the user expected from everyday life. Furthermore, it was expected that the A condition would be judged as very unrealistic since the competing speakers were completely muted, which does not correspond to the real-world scenario. Although preferences were slightly smaller in the A condition than the SF and P conditions, this difference was small and non-significant. Similarly, there were no differences in realism among the different dynamic conditions. Thus, it appears that a user's sense of realism can be maintained, despite fairly unrealistic GCD audio enhancement.

6. EXPERIMENT 3

The previous two experiments looked at different gaze-contingent audio modulation techniques in terms of user preference and performance. However, these experiments did not address the primary question, whether there is a need for these modulation techniques to be gaze-contingent, given that there are other input devices, such as pointers, mice, or joysticks readily available on the market. For immersive virtual reality scenarios a non-contact, hands-free interaction technique is preferred to minimize encumbrance and to maintain immersion and presence. However, to compare the system with high precision pointing devices typically used in desktop and hand-held interfaces, we chose to compare user preference and performance with the gaze-contingent system to 3D pointer (a hand-held portable mouse used for presentations). We expected that participant would be able to perform the speaker identification task equally well with the two input modalities, but the eye-tracker would score higher on the user questionnaire regarding higher realism and immersion.

6.1. Methods

6.1.1. Task. The speech discrimination task was the same as in Experiment 1. As before, users had to listen to a short discussion session and identify the speaker who was discussing the identified topic. Once again trial duration was unlimited, and there were six trials in each block. During each block, the user was asked to either use an eye-tracker or a pointing device as an input device. Each user participated in four blocks of trials in a within-subjects design, two blocks for each interaction device. The order of blocks was counterbalanced across participants in a ABBA or BAAB pattern. For example, if a participant was randomly assigned the eye-tracker as an input device for a first block, then she/he would use a pointing device in the second and third block, and the pointing device again for the fourth block. Following the first and second pair of blocks, the third and fourth pair of blocks users were asked to answer a short questionnaire that compared the two input techniques. Thus, she/he had to fill in two sets of questionnaires during the experiment. The 14 questions on the questionnaire are listed in Table VII.

6.1.2. Procedure. As in the previous two experiments, each session started with users completing a short demographic questionnaire, which was followed by an online audio test and a short English comprehension test. After completing the tests, users were shown a demo trial and were given instruction about how to respond during the trial

Table VII. List of Questions in the Post-block Questionnaire

Q1.	I found it easier to forget that I am listening to virtual speakers rather than real people in ... (Eye-tracker block/Track pad block/ No preference)
Q2.	I found it easier to forget that I was watching a display in ... (Eye-tracker block/Track pad block/ No preference)
Q3.	I felt as if the speakers and I were in different places rather than the same room in ... (Eye-tracker block/Track pad block/ No preference)
Q4.	I felt more aware of being in the real world in ... (Eye-tracker block/Track pad block/ No preference)
Q5.	I felt I knew better what was going to happen next (in terms of sound) in ... (Eye-tracker block/Track pad block/ No preference)
Q6.	Speakers paid closer attention to me in ... (Eye-tracker block/Track pad block/ No preference)
Q7.	The speakers were affected more by who I paid attention to in ... (Eye-tracker block/Track pad block/ No preference)
Q8.	It was easier to adapt to ... (Eye-tracker block/Track pad block/ No preference)
Q9.	The task I had to perform was more difficult in ... (Eye-tracker block/Track pad block/ No preference)
Q10.	I enjoyed more listening to ... (Eye-tracker block/Track pad block/ No preference)
Q11.	Which block was more natural? (Eye-tracker block/Track pad block/ No preference)
Q12.	I felt that I was faster at performing the task with ... (Eye-tracker block/Track pad block/ No preference)
Q13.	I felt that I performed better in the ... (Eye-tracker block/Track pad block/ No preference)
Q14.	I felt more control in the ... (Eye-tracker block/Track pad block/ No preference)

and they were given six trials to practice. This was followed by the experimental blocks. At the end of the experiment, users were debriefed, and all of their questions were answered as well as any additional comments about the experiment recorded.

6.1.3. Stimuli. We used the same set of animated avatars and spatial layout as in Experiment 1. As before, there were three, five, or seven different concurrent speakers. We also used the same corpus of audio clips. However, we used the FF modulation technique for all blocks as it was a consistently an effective technique in the previous experiments.

6.1.4. Participants. Once again, we recruited users from an Undergraduate Research Participant Pool. We recruited 15 users, but 6 had difficulty in calibration, so 9 users were included (6 females and 3 male, ranging in age from 18 to 35). Their English proficiency scores ranged from 90.00% to 100.00%, with a mean a score of 98.88%. All users had uncorrected distance visual acuity of 20/30 or better, with good hearing in both ears. Written informed consent was obtained from all users in accordance with a protocol approved by the York University Ethics Board.

6.2. Results

Consistent with Experiment 1, due to unlimited task execution time, there were very few errors in detecting the correct speaker in both conditions. Specifically, the percent correct for the Eye-tracking condition was $95.37 \pm 9.31\%$, and for the pointing device condition was $94.44 \pm 9.62\%$. In terms of the time, it took to detect the correct speaker, there were no significant differences between the two conditions (Eye-tracker: 1.54 ± 0.75 seconds; Track-pad: 1.36 ± 0.85 seconds; $p = 0.10$). This is despite the fact that all of our users had significant and frequent experience with mice and other computer pointing devices, but none had any experience with eye-tracking and gaze-contingent applications before this experiment.

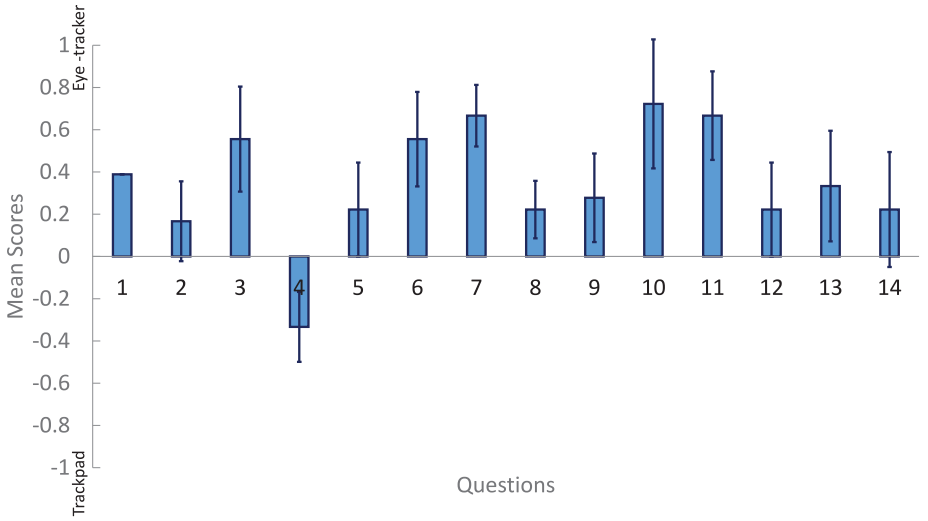

Fig. 15. Mean score for each question. The positive y value indicates an eye-tracker and the negative indicates the track-pad. The error bars represent the standard error of the mean for each question.

Figure 15 shows the mean opinion score for all questions as listed in Table VII. To calculate the final scores, each response was first converted to a numerical score (Eye-tracker: 1; Track-pad: −1; No preference 0) and averaged over the two repetitions of the questionnaire for each subject. The responses for Q3, Q4, and Q9 were negated before plotting because these questions were negatively worded and thus responses indicated a preference for the device that that was not indicated in the response. The plot shows the mean scores across all participants. From the plots, one can see that the highest scores for the eye-tracker were for Q11 (Which block was more natural?), Q10 (I enjoyed more listening to . . .), Q7 (The speakers were affected more by who I paid attention to in . . .), and Q6 (Speakers paid closer attention to me in . . .). The question where participants were most likely to indicate the track-pad was for question Q3 (I felt as if the speakers and I were in different places rather than the same room in . . .), which is a negatively worded question. For successful VR, the users should feel in the same space as the avatars, and thus, these responses indicate a benefit of eye-tracker interaction with the virtual world. Q4 (I felt more aware of being in the real world in . . .) was the only question that indicated a preference for the track-pad and suggested an increased awareness of the real world in the eye-tracking case. This is likely due to the constraints on head motion due to the field of view of the eye-tracker camera. We believe this is not a fundamental limitation as head-mounted eye-trackers are available, and fixed remote eye-trackers are available or coming on the market that have a greatly increased 'head box' volume. For example, we have previously used head-mounted eye-trackers to permit free head-tracking in a six-sided VR immersive environment [Huang et al. 2004]. The technique described in this article is easily extensible to such a scenario.

Finally, it is possible to classify some questions as 'neutral', as the mean scores were not significantly different from '0'. These questions were Q2 (I found it easier to forget that I was watching a display in . . .), (It was easier to adapt to . . .), Q9 (The task I had to perform was more difficult in . . .), Q12 (I felt that I was faster at performing the task with . . .), Q13 (I felt that I performed better in the . . .), and Q14 (I felt more control in the . . .). All these questions reflect the participant's perception of their performance during the task. The 'neutral' responses to these questions are in accordance with their

objective performance during the task – people performed equally well with either of the two input devices.

Both objective and subjective measures showed that people could use both input devices with comparable effectiveness. This is a very promising result as it shows that when used correctly, gaze indeed can be a reliable means of interaction with the VE. The system presented in this article is intended for hands-free use in an immersive VE. As such, it is important to note that the questions related to immersion and enjoyment showed a clear preference for the eye-tracker input. This was despite responses in all conditions being keyed by the user on a keyboard, which would be expected to generally reduce immersion. Hence, we believe this finding supports our proposition that eye-tracking technology and interfaces based on natural gaze behaviour rather than explicit interaction can make virtual simulations more realistic and immersive.

7. GENERAL DISCUSSION

In the three studies presented above, we examined several factors such as multiple virtual speakers and various sound modulation techniques that could be used to create realistic VR social interactions in simulations and games. The next subsections will discuss the fundamental differences between the real-world scenarios and simulated VR environments and how the proposed GCD techniques mitigated the differences. This discussion will follow with a discussion of generalization to other tasks, the impact and the role of gaze-contingent applications, and how our results can impact the next generation of VR development.

7.1. Real World Versus Virtual Reality

When designing realistic VR, it is important to understand what cues or combination of cues are missing in VR versus the real world and try to compensate for these cues in the simulated environments. One specific example of such discrepancy is that the rich spatial audio cues underlying the impressive human capabilities demonstrated in the CP effect are not normally supported in current VR. The reasons for not incorporating these spatial audio cues may partly be the lack of appreciation of the effect and its benefits but the cost, lack of data for user-specific filters, demands on spatial tracking, and limited computational resources are also limiting factors. However, when such effects are not supported in VR implementation, it can directly impact the ability to interact in group social scenarios and other cluttered auditory environments.

There are many occasions similar to a CP scenario; however, simulating the full range of spatial sound cues present in a realistic CP is very computationally intensive, requires obtaining individualized HRTF, and is not always possible; hence, simplifications are necessary. Furthermore, many applications have limited range of variation and it has been established that similar voices impair content discrimination [Brungart et al. 2001; Egan et al. 1954]. Indeed, in our experiment, users had a hard time discriminating between speakers without any audio augmentation. However, with augmentation, performance approached expectations from real-world experience. This behavioural improvement is consistent with Brungart and Simpson [2007], who observed improvements in speaker segregation by amplifying the volume of a speaker and with Best et al. [2007], who showed that spatial knowledge of where the object is located can help in enhancing source identification. Our results are also consistent with studies that looked at the effect of gaze direction on sound source localization [Lewald and Ehrenstein 1996; Lewald 1998; Maddox et al. 2014].

7.2. Experimental Task

The ability to simulate multiple avatars is important for studying social interaction mechanisms [Wilms et al. 2010] or when building games [Cullen et al. 2012].

Furthermore, understanding the nature of interaction one might have with the avatars in VR will impact the choice of modulation techniques and subsequently users' performance and experience. For all three experiments presented in this article, the task was to find a specific speaker among a group of speakers who is speaking about an assigned topic. In other words, the selected task was focussed on the acquisition of the desired information with an intention to achieve highest performance (either by finding the desired speaker as fast as possible or as accurately as possible). It is possible people preferred techniques that supported best performance with the given task. For example, this might explain why the A technique appears to be superior and is judged as natural, despite it being objectively unnatural. However, we acknowledge that this is not the only possible task. For completeness, future studies should also look at cases where relevant information has to be collected from multiple speakers and/or conversations, which is common in social scenarios. It would be important to know whether the preferences amongst the different techniques would vary with task.

In addition to selecting an appropriate task, it important to pay attention to other details that might impact users' performance. For instance, it was important to make sure that avatars were animated and their facial animation was synchronized with the text that they were saying. On many occasions during the debriefing session, users reported that they used animation for verification and as an additional cue to help them to successfully complete an identification task. On the other hand, we could not control shading of the faces of our avatars as their facial textures were pre-made prior to being placed in the experimental scene. While this was not commented on during debriefing, it is possible it had some impact on users' sense of immersion.

Speaker order was not very informative in our experiment as all speakers spoke simultaneously. Temporal information, according to Best et al. [2007], can help in sound discrimination when targets are extremely similar or unfamiliar. Our speakers were purposely selected to be similar to each other (all male voices), and hence gaze-contingent techniques with instant or fast audio modulations were the most accurate as users could associate the source with their actions through the close temporal coupling. Furthermore, directing attention to a particular sensory channel not only helps in localizing and discriminating between sound sources, but also helps to overcome errors or imprecision in the other modality [Rimell and Owen 2000]. Hence, we expect that since visual attention was involved, people could better tolerate some audio inconsistencies, such as slow speaker modulation. Conversely, when attending to audio channels they likely could better tolerate any imprecision associated with visual display and gaze-tracking. It is possible the effect of such factors could be reflected in workload but not performance, and future studies could use some measures of workload to verify whether the GCD techniques affected workload.

Finally, all three experiments dealt with a static scene, which provided experimental control but reduced the generality of the scenario. In a future, dynamic experiments should also be conducted to utilize the spatial capabilities of the proposed techniques to the fullest.

7.3. GCD Impact

In the context of a spatial sound segregation and verification, gaze-contingent systems are potentially very advantageous, as the user always knows which speaker they are looking at each instant and can verify what they hear is consistent with what they see by observing the speaker's lip animation. This of course can be achieved by other input devices, however, for immersive VR, the primary advantages of gaze-contingent display is the natural, unobtrusive nature of the interaction and the lack of requirement for interaction with the hands or voice or other deliberate action. In addition, the use of gaze-tracking in this article provided similar performance to mouse and other selection

techniques but is natural and frees the hands for VR interaction (through gesture or input devices) [Argelaguet and Andujar 2013] and the voice for social interaction. Although eye-tracking techniques are not readily available to all users, costs are rapidly decreasing and some eye-tracking is available for current generation head-mounted displays, such as the Oculus Rift and HTC Vive [Sridharan et al. 2015; Sidorakis et al. 2015; Orlosky et al. 2015]. Therefore, these techniques are very relevant for HMD-based VR as well as immersive projection technology-based VR.

Nonetheless, most VR and AR head-sets incorporate head-tracking only. Our users typically made small head movements consistent with the relatively central location of the avatars, and thus head base selection would not provide for discrimination of the different speakers. Head-tracking is of course critical for accurate HRTF simulation but also for lower fidelity spatialized sound simulation, such as binaural sound. Our GCD techniques using both head- and eye-tracking could be combined with head-tracking for spatialized sound (perhaps of lower fidelity than individualized HRTFs). In the future, it would be good to compare the effectiveness of head-contingent audio highlighting with gaze-contingent audio highlighting and also to explore the combination of gaze-contingent audio highlighting with modest fidelity spatialized sound.

7.4. Future Development and Impact

The audio highlighting described in this article could be one technique in a suite of gaze-contingent techniques for VR applications. In addition, although we think that our audio techniques can be best used for VR applications, they can also be incorporated for multi-person video conferences applications well. Especially, this can be beneficial for applications that can use gaze-information as a secondary source of information to mark or identify common areas of interest between participants or to orchestrate multiple conversations within the same auditory space.

8. CONCLUSIONS

In this article, we have demonstrated the use of gaze-contingent audio display to augment user experience in social VEs. Specifically, we looked at how background modulation as well as a change in volume for the speaker of interest improved speaker identification as well as how these techniques are rated by the users along different aspects of realism and immersion. Our work showed that rapid enhancement of the sound of the source one is attending to (regardless of the speed of the background modulation) can both improve system realism and simultaneously improve the quality and speed of auditory information acquisition.

We also compared eye-tracking versus a pointing device as an input device for VR immersive application. In our scenario, gaze is naturally directed towards conversational partners, and gaze is thus an unobtrusive measure of attentional selection. Such use of natural behaviours and hands-free interaction can promote immersion in a VE, since it does not require breaking out of the simulated world, and furthermore, it frees the hands for direct interaction with objects in the VE. In agreement with this proposition, we found that people were able to perform equally well with both input devices but felt more immersed and enjoyed the eye-tracker better than with a traditional pointing device.

This article has demonstrated how gaze-contingent audio modulation techniques can improve user experience and task performance in an important scenario. However, we believe the potential for GCD audio is broader and that the presented techniques can significantly improve user experience for a wide range of use cases. These techniques are natural fits for VR and augmented reality (AR) domains in particular because they leverage and utilize natural and effectively effortless gaze behaviours.

In the future, we would like to look more closely on comparing our sound modulation techniques with high-fidelity, individually tailored spatially localized sound. We hypothesize that estimates of user gaze can be used to reduce the requirements for audio fidelity of unattended objects, and therefore, this approach has the potential to reduce the computational cost of high-fidelity audio rendering.

ACKNOWLEDGMENTS

The authors would give a special thank you to Carly Hylton and Anna Kiseleva for their help in setting up and conducting experimental sessions. We would like to thank Katherine Allison for the voiceover in the supplementary video.

REFERENCES

F. Argelaguet and C. Andujar. 2013. A survey of 3D object selection techniques for virtual environments. *Computers & Graphics* 37, 3 (2013), 121–136.

J. J. Baldis. 2001. Effects of spatial audio on memory, comprehension, and preference during desktop conferences. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'01)*. ACM, New York, NY, 166–173.

D. W. Batteau. 1967. The role of the pinna in human localization. *Royal Society of London B: Biological Sciences* 168 (1967), 158–180.

S. Benford, C. Greenhalgh, and D. Lloyd. 1997. Crowded collaborative virtual environments. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*. 7, 2 (1997), 59–66.

G. Bente, F. Eschenburg, and N. C. Krämer. 2007. Virtual gaze. A pilot study on the effects of computer simulated gaze in avatar-based conversations. *Virtual Reality* 4563 (2007), 185–194.

V. Best, E. J. Ozmeral, and B. G. Shinn-Cunningham. 2007. Visually-guided attention enhances target identification in a complex auditory scene. *Journal for the Association for Research in Otolaryngology* 8 (2007), 294–304.

M. Billinghurst, J. Bowskill, M. Jessop, and J. Morphett. 1998. A wearable spatial conferencing space. In *Proceedings of the 2nd International Symposium on Wearable Computers, 1998*. IEEE, 76–83.

F. Biocca, C. Harms, and J. Burgoon. 2003. Toward a more robust theory and measure of social presence: Review and suggested criteria. *Presence* 12 (2003), 456–480.

R. A. Bolt. 1981. Gaze-orchestrated dynamic windows. In *SIGGRAPH Computer Graphics*. ACM, 109–119.

D. S. Brungart and B. D. Simpson. 2005. Optimizing the spatial configuration of a seven-talker speech display. *ACM Transactions on Applied Perception* 2, 4 (Oct. 2005), 430–436.

D. S Brungart and B. D. Simpson. 2007. Cocktail party listening in a dynamic multitalker environment. *Perception and Psychophysics* 69 (2007), 79–91.

D. S. Brungart, B. D. Simpson, M. A. Ericson, and K. R. Scott. 2001. Informational and energetic masking effects in the perception of multiple simultaneous talkers. *The Journal of the Acoustical Society of America* 110 (2001), 2527–2538.

E. Castellina and F. Corno. 2008. Multimodal gaze interaction in 3D virtual environments. In *COGAIN 2008 Communication, Environment and Mobility Control by Gaze* (2008). 1–5.

E. C. Cherry. 1953. Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America*. 25 (1953), 975–979.

B. Cullen, D. Galperin, K. Collins, B. Kapralos, and A. Hogue. 2012. The effects of audio on depth perception in S3D games. In *Proceedings of the Audio Mostly Conference: A Conference on Interaction with Sound*. ACM, 32–39.

S. Deo, M. Billinghurst, N. Adams, and J. Lehikoinen. 2007. Experiments in spatial mobile audio-conferencing. In *Proceedings of the 4th International Conference on Mobile Technology, Applications, and Systems and the 1st International Symposium on Computer Human Interaction in Mobile Technology (Mobility'07)*. ACM, New York, NY, 447–451.

R. Drullman and A. W. Bronkhorst. 2000. Multichannel speech intelligibility and talker recognition using monaural, binaural, and three-dimensional auditory presentation. *Journal of the Acoustical Society of America* 107, 4 (2000), 2224–2235.

A. T. Duchowski. 2000. Acuity-matching resolution degradation through wavelet coefficient scaling. *Image Processing* 9 (2000), 1437–1440.

A. T. Duchowski. 2007. *Eye Tracking Methodology*. Springer Science & Business Media.

J. P. Egan, E. C. Carterette, and E. J. Thwing. 1954. Some factors affecting multi-channel listening. *Journal of the Acoustical Society of America* 26, 5 (1954), 774–782.

M. R. Frater, J. F. Arnold, and A. Vahedian. 2001. Impact of audio on subjective assessment of video quality in videoconferencing applications. *Circuits and Systems for Video Technology* 11 (2001), 1059–1062.

W. S. Geisler and J. S. Perry. 1998. A real-time foveated multiresolution system for low-bandwidth video communication. In *Human Vision and Electronic Imaging III*, Vol. 3299. International Society for Optics and Photonics, 294–305.

S. Goose, J. Riedlinger, and S. Kodlahalli. 2005. Conferencing3: 3D audio conferencing and archiving services for handheld wireless devices. *International Journal of Wireless and Mobile Computing* 1, 1 (Nov. 2005), 5–13.

D. Hindus, M. S. Ackerman, S. Mainwaring, and B. Starr. 1996. Thunderwire: A field study of an audio-only media space. In *Proceedings of the 1996 ACM Conference on Computer Supported Cooperative Work (CSCW'96)*. ACM, New York, NY, 238–247.

S. Ho, T. Foulsham, and A. Kingstone. 2015. Speaking and listening with the eyes: Gaze signaling during dyadic interactions. *PLoS ONE* 10, 8 (2015), 1–18.

H. Huang, R. S. Allison, and M. Jenkin. 2004. Combined head-eye tracking for immersive virtual reality. In *Proceedings of the 2004 14th International Conference on Artificial Reality and Telexistance.*

A. Hyrskykari, P. Majaranta, and K. J. Räihä. 2005. From gaze control to attentive interfaces. In *Proceedings of the International Conference on Human-Computer Interaction*.

R. J. Jacob. 1991. The use of eye movements in human-computer interaction techniques: What you look at is what you get. *Information Systems* 9 (1991), 152–169.

R. Kilgore, M. Chignell, and P. Smith. 2003. Spatialized audioconferencing: What are the benefits?. In *Proceedings of the 2003 Conference of the Centre for Advanced Studies on Collaborative Research (CASCON'03)*. IBM Press, 135–144.

F. Kistler, M. Wißner, and E. André. 2010. Level of detail based behavior control for virtual characters. In *Intelligent Virtual Agents*. Springer-Verlag, 118–124.

P. Kortum and W. S. Geisler. 1996. Implementation of a foveated image coding system for image bandwidth reduction. *Human Vision and Electronic Imaging* 2657 (1996), 350–360.

B. Kunka and B. Kostek. 2010. Exploiting audio-visual correlation by means of gaze tracking. *International Journal of Computer Science and Applications* 7 (2010), 104–123.

B. Kunka, B. Kostek, M. Kulesza, P. Szczuko, and A. Czyzewski. 2010. Gaze-tracking-based audio-visual correlation analysis employing quality of experience methodology. *Intelligent Decision Technologies* 4 (2010), 217–227.

J. Lewald. 1998. The effect of gaze eccentricity on perceived sound direction and its relation to visual localization. *Hearing Research* 115 (1998), 206–216.

J. Lewald and W. H. Ehrenstein. 1996. The effect of eye position on auditory lateralization. *Experimental Brain Research* 108 (1996), 473–485.

V. Losing, T. Pfeiffer, L. Rottkamp, and M. Zeunert. 2014. Guiding visual search tasks using gaze-contingent auditory feedback. In *Proceedings of the International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*. ACM, 1093–1102.

R. K. Maddox, D. A. Pospisil, G. C. Stecker, and A. K. Lee. 2014. Directing eye gaze enhances auditory spatial cue discrimination. *Current Biology* 24 (2014), 748–752.

G. Marmitt and A. T. Duchowski. 2002. Modeling visual attention in VR: Measuring the accuracy of predicted scanpaths. *Eurographics*. 217–226.

G. Mastoropoulou, K. Debattista, A. Chalmers, and T. Troscianko. 2005. The influence of sound effects on the perceived smoothness of rendered animations. In *Proceedings of the Symposium on Applied Perception in Graphics and Visualization*. ACM, 9–15.

N. Megiddo. 2003. System and methodology for video conferencing and internet chatting in a cocktail party style. (May 6, 2003). US Patent 6,559,863. Filing date: Feb. 11, 2000.

G. A. Miller. 1947. The masking of speech. *Psychological Bulletin* 44 (1947), 105–129.

A. Mortlock, D. Machin, S. McConnell, and P. Sheppard. 1997. Virtual conferencing. *BT Technology Journal* 15, 4 (1997), 120–129.

H. Nakanishi, C. Yoshida, T. Nishimura, and T. Ishida. 1996. FreeWalk: Supporting casual meetings in a network. In *Proceedings of the 1996 ACM Conference on Computer Supported Cooperative Work (CSCW'96)*. ACM, New York, NY, 308–314.

J. O'Donovan, J. Ward, S. Hodgins, and V. Sundstedt. 2009. Rabbit run: Gaze and voice based game interaction. In *Proceedings of the Eurographics Ireland Workshop.*

K. Okada, F. Maeda, Y. Ichikawaa, and Y. Matsushita. 1994. Multiparty videoconferencing at virtual social distance: MAJIC design. In *Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work (CSCW'94)*. ACM, New York, NY, 385–393.

J. Orlosky, T. Toyama, K. Kiyokawa, and D. Sonntag. 2015. ModulAR: Eye-controlled vision augmentations for head mounted displays. *IEEE Transactions on Visualization and Computer Graphics* 21, 11 (2015), 1259–1268.

D. J. Parkhurst and E. Niebur. 2002. Variable-resolution displays: A theoretical, practical, and behavioral evaluation. In *Human Factors: The Journal of the Human Factors and Ergonomics Society* 44 (2002), 611–629.

N. Pelechano, C. Stocker, J. Allbeck, and N. Badler. 2008. Being a part of the crowd: Towards validating VR crowds using presence. In *Proceedings of the International Joint Conference on Autonomous Agents and Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 136–142.

A. N. Rimell, N. J. Mansfield, and D. Hands. 2008. The influence of content, task and sensory interaction on multimedia quality perception. *Ergonomics* 51 (2008), 85–97.

A. N. Rimell and A. Owen. 2000. The effect of focused attention on audio-visual quality perception with applications in multi-modal codec design. In *Proceedings of the Conference on Acoustics, Speech, and Signal Processing*. IEEE, 2377–2380.

A. Sellen, B. Buxton, and J. Arnott. 1992. Using spatial cues to improve videoconferencing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'92)*. ACM, New York, NY, 651–652.

A. J. Sellen. 1992. Speech patterns in video-mediated conversations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'92)*. ACM, New York, NY, 49–59.

S. A. Shamma, M. Elhilali, and C. Micheyl. 2011. Temporal coherence and attention in auditory scene analysis. *Trends in Neurosciences* 34 (2011), 114–23.

N. Sidorakis, G. A. Koulieris, and K. Mania. 2015. Binocular eye-tracking for the control of a 3D immersive multimedia user interface. In *Proceedings of the 2015 IEEE 1st Workshop on Everyday Virtual Reality (WEVR)*. 15–18.

S. Sridharan, J. Pieszala, and R. Bailey. 2015. Depth-based subtle gaze guidance in virtual reality environments. In *Proceedings of the ACM SIGGRAPH Symposium on Applied Perception (SAP'15)*. ACM, New York, NY, 132–132.

I. Starker and R. A Bolt. 1990. A gaze-responsive self-disclosing display. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 3–10.

L. B. Stelmach and W. J. Tam. 1994. Processing image sequences based on eye movements. In *International Symposium on Electronic Imaging: Science and Technology*, Vol. 2179. International Society for Optics and Photonics, SPIE, 90–98.

S. S. Stevens. 1955. The measurement of loudness. *The Journal of the Acoustical Society of America* 27 (1955), 815–829.

J. W. Strutt. 1907. On our perception of sound direction. *Philosophical Magazine* 13 (1907), 214–232.

V. Tanriverdi and R. J. K. Jacob. 2000. Interacting with eye movements in virtual environments. In *Human Factors in Computing Systems*. ACM, 265–272.

N. Tsumura, C. Endo, H. Haneishi, and Y. Miyake. 1996. Image compression and decompression based on gazing area. *Human Vision and Electronic Imaging* 2657 (1996), 361–367.

H. Turner and D. Firth. 2007. *Generalized Nonlinear Models in R: An Overview of the Gnm Package. (R package version 1.0-6)*. Technical Report. 472.

H. Turner and D. Firth. 2012. Bradley-Terry Models in R: The BradleyTerry2 Package. *Journal of Statistical Software* 48, 9 (2012).

L. Twardon, H. Koesling, A. Finke, and H. Ritter. 2013. Gaze-contingent audio-visual substitution for the blind and visually impaired. In *Proceedings of the International Conference on Pervasive Computing Technologies for Healthcare*. Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, 129–136.

J. van der Kamp and V. Sundstedt. 2011. Gaze and voice controlled drawing. In *Proceedings of the 1st Conference on Novel Gaze-Controlled Applications (NGCA'11)*. ACM, New York, NY.

R. Vertegaal. 1999. Designing awareness with attention-based groupware. In *INTERACT*. 245–255.

M. Vidal, R. Bismuth, A. Bulling, and H. Gellersen. 2015. The royal corgi: Exploring social gaze interaction for immersive gameplay. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 115–124.

M. Vinnikov and R. S. Allison. 2013. Gaze-contingent simulations of visual defects in virtual environment: Challenges and limitations. In *Proceedings of the CHI 2013 Workshop on "Gaze Interaction in the Post–WIMP World."*

M. L. H. Võ, T. J. Smith, P. K. Mital, and J. M. Henderson. 2012. Do the eyes really have it? Dynamic allocation of attention when viewing moving faces. *Journal of Vision* 12 (2012), 1–14.

M. Wilms, L. Schilbach, U. Pfeiffer, G. Bente, G. R. Fink, and K. Vogeley. 2010. It's in your eyes–using gaze-contingent stimuli to create truly interactive paradigms for social cognitive and affective neuroscience *Social Cognitive and Affective Neuroscience* 5 (2010), 98–107.

S. Winkler and C. Faller. 2005. Audiovisual quality evaluation of low-bitrate video. In *Human Vision and Electronic Imaging X*, Bernice E. Rogowitz, Thrasyvoulos N. Pappas, and Scott J. Daly (Eds.), Vol. 5666. SPIE, 139–148.

H. A. Witkin and T. Leventhal. 1952. Sound localization with conflicting visual and auditory cues. *Journal of Experimental Psychology* 43 (1952), 58–67.

T. Yonezawa, H. Yamazoe, A. Utsumi, and S. Abe. 2007. Gaze-communicative behavior of stuffed-toy robot with joint attention and eye contact based on ambient gaze-tracking. In *Proceedings of the International Conference on Multimodal Interfaces.* 140–145.

Received August 2015; revised December 2016; accepted December 2016