Spatial Misregistration of Virtual Human Audio: Implications of the Precedence Effect

David M. Krum, Evan A. Suma, and Mark Bolas

Institute for Creative Technologies
12015 Waterfront Drive, Playa Vista, CA 90094, USA
{krum,suma,bolas}@ict.usc.edu

Abstract. Virtual humans are often presented as mixed reality characters projected onto screens that are blended into a physical setting. Stereo loudspeakers to the left and right of the screen are typically used for virtual human audio. Unfortunately, stereo loudspeakers can produce an effect known as precedence, which causes users standing close to a particular loudspeaker to perceive a collapse of the stereo sound to that singular loudspeaker. We studied if this effect might degrade the presentation of a virtual character, or if this would be prevented by the ventriloquism effect. Our results demonstrate that from viewing distances common to virtual human scenarios, a movement equivalent to a single stride can induce a stereo collapse, creating conflicting perceived locations of the virtual human's voice. Users also expressed a preference for a sound source collocated with the virtual human's mouth rather than a stereo pair. These results provide several design implications for virtual human display systems.

Keywords: virtual human audio, spatial sound, stereo audio, precedence effect, ventriloquism effect, mixed reality.

1 Introduction

Computer controlled virtual humans are an increasingly important part of applications in entertainment, training, therapy, novel human-computer interfaces, and social research. Often a 3D mixed reality presentation is preferred, where life-sized virtual characters are blended into a staged physical setting. Digital projectors are often the display of choice, since they are relatively inexpensive, can be used without head tracking, do not require users to don any display gear, and can be seen by multiple users at a number of angles and distances.

While a single loudspeaker located near a character's mouth and chest can portray the voice of a single virtual human character, this placement can be problematic. With a rear projected screen configuration, the loudspeaker would likely block the video image. Furthermore, placing the loudspeaker behind the screen would result in muffled audio. While there are perforated screens that allow sounds to pass, these screens require front projection. Rear projection is more desirable since it prevents users from accidentally blocking the projection

Y. Nakano et al. (Eds.): IVA 2012, LNAI 7502, pp. 139–145, 2012.

and casting shadows across the character. A rear projection display combined with stereo loudspeakers is thus a common compromise in many installations.

As with any stereo pair, this configuration is subject to the precedence effect [15,19], which can interfere with stereo spatialization. The interference occurs when a listener is standing much closer to one of the two loudspeakers in a stereo pair. At this location, the wavefront from the nearby loudspeaker arrives sooner than, or precedes, the other loudspeaker's wavefront. The human perceptual system has an echo cancellation process, which ignores the second wavefront. Only the initial wavefront is perceived, causing the perceived sound location to collapse to the nearby loudspeaker, breaking the stereo spatialization. Our concern was that the precedence effect might cause a virtual human's voice to shift to the left or right as the listener moved around. Adding to our uncertainty was a second phenomenon, the ventriloquism effect, which might counteract the precedence effect. The ventriloquism effect can create the perception that a voice or sound, generated elsewhere, is emanating from the visual image of a temporally related source [5,7,13,18]. A previous study of interactions between ventriloquism and precedence effects demonstrated that they can work in concert, i.e. strengthening the perceived locality of a sound with a visual image that is coincident with a preceding sound source [11]. However, that study did not examine how the two effects might work in opposition.

With the goal of greater versimilitude in mixed reality training, it is problematic if a character's voice emanates from a point that is perceptibly offset from the character's mouth and body. Furthermore, a breakdown in spatialization can have negative effects on conversational interactions. Studies have shown that spatialization of multiple voices can increase speech comprehension, voice identification, and understanding [9,12,2].

Our goal was to examine the impact of precedence effect in a mixed reality virtual human presentation. Would it negatively impact a user's perception of the virtual human, or would it be masked by the ventriloquism effect working in opposition?

2 Related Work

A number of researchers believe that spatialized audio is an important sensory cue and have worked to improve 3D spatialized audio for users of mixed, augmented, and virtual reality [14,16,17,8]. Beyond stereophonic sound and its variants, current spatial audio reproduction systems include headphone based techniques using binaural audio and head related transfer functions [6] as well as techniques using arrays of loudspeakers like Ambisonics [10] and wavefield synthesis [3,4]. Headphone based techniques are less appealing since users must wear an additional device, preventing a simple "walk up and interact" experience. Wavefield synthesis requires large numbers of loudspeakers, perhaps hundreds or more, increasing cost and complexity. The Ambisonic technique typically requires four or more loudspeakers, as well as decoding hardware, and can have sound reproduction issues in large spaces without sound treatment to control echo and

reverberation. Furthermore, wavefield synthesis and Ambisonic techniques may be unnecessary for virtual humans displayed on projected screens. The virtual characters appear on a screen in front of users, so the ability to present spatial sound from any point surrounding the users is simply unnecessary.

3 Methods and Apparatus

To determine if the precedence effect can alter the perception of a projected virtual human, we designed and conducted a mixed design study where participants listened to a virtual human reading literary passages. Participants were divided into three equal size groups for placement at one of three physical locations in front of the virtual human, representing the between subjects condition. Audio presentation was the within subjects condition.

Thirty-six participants, over the age of 18, with 20/20 corrected vision and self-identifying as having hearing in both ears were enrolled through email and the Craigslist website. The gender ratio was evenly balanced and participants ranged from 20 to 67 years of age (M=37.6, SD=13.6).

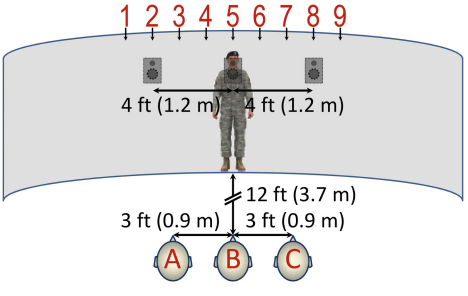

Fig. 1. Three loudspeakers, located behind a perforated (sound transparent) screen, provided audio to the left of the projected virtual human, at the center, or in stereo. Participants stood at positions: A, B, or C.

Participants were placed at one of three positions approximately 12 feet (3.7 m) in front of the virtual character. Position A was 3 feet (0.9 m) to the left of the screen mid-line, position B was on the mid-line, and position C was 3 feet (0.9 m) to the right of the mid-line (see Figure 1). This lateral distance was chosen, based on our experience with mixed reality virtual human installations for museums and military training, to represent the distance of a single stride that a single user might make or a comfortable distance between two participants.

The virtual human used throughout the study was a male soldier using a Cepstal LLC text-to-speech voice. The character was rendered using the Panda3D graphics library and animated to provide appropriate visemes, eye blinks, and breathing motions. Passages were selected from Herman Melville's *Moby Dick* for variety, duration, and good delivery by the character.

The character was projected onto a curved screen approximately 8.5 feet (2.6 m) tall and 31.3 feet (9.5 m) in width along the curve (see Figure 1). Three loudspeakers were placed behind the screen at the center, 4 feet (1.2 m) to the left of center, and 4 feet (1.2 m) to the right of center. The loudspeakers were Mackie HR824 high resolution studio monitors, mounted 57 inches (145 cm) high to bring them to the height of the virtual character's mouth and chest. A perforated screen was used, which allowed sound be heard through the screen. As previously mentioned, this perforated screen requires front projection and is thus not optimal for normal interaction with virtual characters. For this study, the perforated screen was satisfactory as user movement was restricted. Sound pressure levels were calibrated to provide matched values between audio presentations. Loudspeakers were also swapped halfway through the study to help counterbalance any tonal differences between loudspeakers. Some sound treatment was also applied behind the loudspeakers and screen to limit reverberations.

In the first phase of the study, each participant listened to three passages read by the virtual human and presented once each by either the left loudspeaker, the center loudspeaker, or a stereo pair. Participants were not told which audio presentation was being used. The order of the audio presentation was fully randomized. After each passage, participants were given two survey questions related to virtual human co-presence, based on the Bailenson et al. social presence questionaire [1]. ("I perceived the virtual human as being only a computerized image, not a real person." and "I perceived that the virtual human was present in the room with me.") Participants were asked to respond on 7 point scales. Participants then indicated the apparent horizontal location of the voice by referencing a set of markers from 1 to 9 placed along the top of the screen. The markers were spaced approximately 16 inches (41 cm) apart, with the 5 marker located at the center, above the center loudspeaker and the character.

In the second phase, participants listened to the virtual human's delivery of four sentence pairs. Each sentence pair consisted of the same sentence, repeated twice, but using different loudspeakers. Two of the sentence pairs were presented first in stereo and then by the center loudspeaker. The other two pairs were presented first by the center loudspeaker and then in stereo. Order of presentation was randomized. Participants were asked, for each sentence pair, "Which line was delivered more like a real person?", and asked to select either the first or second sentence. Participants were not told which audio configuration was used.

4 Results and Discussion

A mixed ANOVA statistical test was performed to determine if the within-subjects condition of audio presentation as well as the between-subjects condition of listener position created significant differences in the perceived location of the sound source at the $\alpha = .05$ level. Since Mauchly's Test of Sphericity indicated a possible violation of sphericity for the within-subject effects of audio presentation, we performed a Greenhouse-Geisser correction.

Data from the first phase of the study is listed in Table 1. A significant main effect of audio presentation (left, center, or stereo) was observed in perceived

location, $F(2, 66) = 32.40, p < .001, \eta_p^2 = .50$. The effect of audio presentation was expected as the sound source changed position between audio presentations. This effect thus helps to confirm the validity of the experimental configuration. An examination of the 95% confidence intervals reveals that the perceived location of speech produced by the left loudspeaker, 95% CI [3.27, 4.45], is well separated and clearly different from the center [5.92, 6.35] and stereo [5.24, 6.09] presentations (see Figure 2a).

Table 1. Perceived Location by Audio Presentation and Listener Position. (Location values signify: 1=Left, 5=Middle, 9=Right.)

Audio Presentation	Listener Position	Mean Perceived Location	Standard Deviation
Left	A:Left	3.42	1.505
	B:Middle	4.33	2.270
	C:Right	3.83	1.267
	Total	3.86	1.726
Center	A:Left	6.67	0.778
	B:Middle	5.83	0.577
	C:Right	5.92	0.515
	Total	6.14	0.723
Stereo	A:Left	4.33	1.155
	B:Middle	5.00	1.279
	C:Right	7.67	1.303
	Total	5.67	1.897

The between subjects variable, listener position (A:Left, B:Middle, or C:Right) was observed to create a significant difference in perceived location $F(2, 33) = 5.56, p = .008, \eta_p^2 = .25$. More importantly, the crossing of the trendlines (see Figure 2b) shows a significant interaction effect for the stereo condition in combination with listener position $F(4, 66) = 10.16, p < .001, \eta_p^2 = .38$. For listeners at the rightmost position, the stereo sound source is perceived at the right side of the screen (larger numbers). For listeners at the leftmost position, the stereo sound source is perceived towards the left side (smaller numbers). This crossover is evidence of the precedence effect occuring in the stereo loudspeaker condition. There does appear a slight systemic shift to the right, perhaps due to room acoustics, as well as some pull towards the virtual human's central visual image, possibly due to the ventriloquism effect. However, the magnitudes of these effects do not obscure the interaction which suggests the precedence effect.

We did not observe any significant effect of audio presentation on the two questions concerning co-presence at the $\alpha = .05$ level with a mixed ANOVA. Several sources of variance may have affected these measures. Many participants may have been unfamiliar with virtual humans and had little common reference for co-presence. We also observed some possible confusion concerning the direction of the scales for the two questions. A scale reversal is present in the original social presence questionnaire from which these questions were adapted. Furthermore, the baseline realism of the virtual human's voice and behavior were limited, possibly overwhelming any contribution of varying audio presentation.

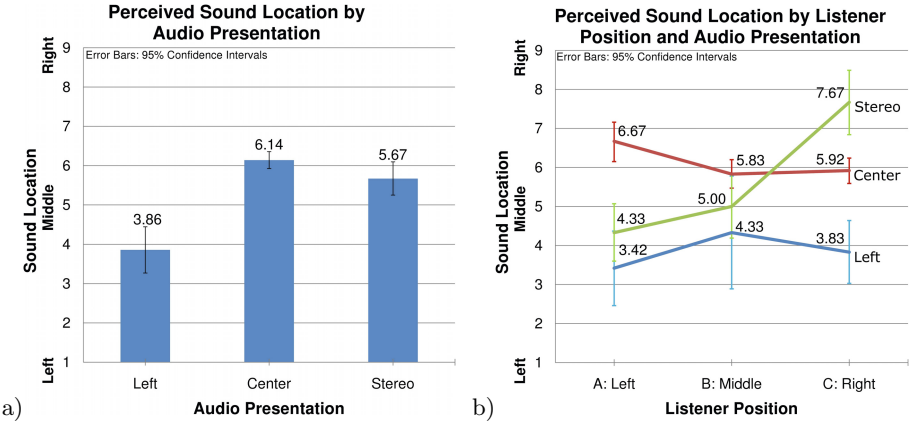

Fig. 2. a) Significant differences in perceived location were observed for audio conditions. Localization of speech from the left loudspeaker clearly differed from center and stereo presentations. b) A significant interaction effect on perceived location was observed for listener position and the stereo audio condition. Listeners on the left localized the stereo sound to the left, while listeners on the right localized it to the right.

For the second phase of the study, the sentence pair trials, loudspeaker preferences for each of the four trials were recoded numerically (0=center, 1=stereo) and then summed to provide an overall preference score. A One-Sample Wilcoxon Signed Ranks Test was conducted to compare these scores against an expected median value of 2.0, corresponding to random chance. The observed median of 1.0 indicated a significant preference for the center speaker location, $Z = -3.71, p < .001$, suggesting that a single loudspeaker delivered better realism.

5 Conclusion and Future Work

This study demonstrates that the precedence effect can occur in a typical stereo configuration for virtual human audio, causing misperception of the audio source. The offset distance tested can easily occur with a single mobile participant or multiple participants. These results demonstrate limitations of stereo loudspeaker pairs in supporting user movement and multiple users around virtual humans. While adjustments to stereo phase and panning can compensate for motion of a single user, user tracking is required, and these adjustments cannot scale to multiple users. Designers should examine the range of motion and number of users required and select complementary audio/visual components that can robustly collocate virtual human audio and visual imagery for the given installation. Consideration of perforated screens and individual loudspeakers assigned to each virtual character may be warranted. We expect that these results will also inform development of new technologies for presenting virtual human audio. To be of interest to virtual human installation designers, these approaches

should be compatible with projected displays and attempt to better replicate the proximal sound field of a human voice.

References

1. Bailenson, J.N., Blascovich, J., Beall, A.C., Loomis, J.M.: Equilibrium theory revisited: Mutual gaze and personal space in virtual environments. Presence-Teleop. Virt. 10, 583–598 (2001)
2. Baldis, J.J.: Effects of spatial audio on memory, comprehension, and preference during desktop conferences. In: ACM CHI, pp. 166–173 (2001)
3. Berkhout, A.J.: A holographic approach to acoustic control. J. Audio Eng. Soc. 36(12), 977–995 (1988)
4. Berkhout, A.J., De Vries, D., Vogel, P.: Acoustic control by wave field synthesis. J. Acoust. Soc. Am. 93, 2764–2778 (1993)
5. Bertelson, P.: Chapter 14 ventriloquism: A case of crossmodal perceptual grouping. In: Gisa Aschersleben, T.B., Msseler, J. (eds.) Cognitive Contributions to the Perception of Spatial and Temporal Events, Advances in Psychology, vol. 129, pp. 347–362. North-Holland (1999)
6. Blauert, J.: Räumliches Hören (Spatial Hearing). S. Hirzel-Verlag, Stuttgart (1974)
7. Choe, C., Welch, R., Gilford, R., Juola, J.: The ventriloquist effect: Visual dominance or response bias? Atten. Percept. Psycho. 18, 55–60 (1975)
8. Courgeon, M., Rebillat, M., Katz, B., Clavel, C., Martin, J.C.: Life-sized audiovisual spatial social scenes with multiple characters: MARC & SMART-I2. In: Meeting of the French Association for Virtual Reality (2010)
9. Ericson, M.A., Brungart, D.S., Simpson, B.D.: Factors that influence intelligibility in multitalker speech displays. Int. J. Aviat. Psychol. 14, 313–334 (2004)
10. Fellget, P.: Ambisonics. part one: General system description. Studio Sound 17, 20–22, 40 (August 1975)
11. Harima, T., Abe, K., Takane, S., Sato, S., Sone, T.: Influence of visual stimulus on the precedence effect in sound localization. Acoust. Sci. Tech. 30(4), 240–248 (2009)
12. Ihlefeld, A., Sarwar, S.J., Shinn-Cunningham, B.G.: Spatial uncertainty reduces the benefit of spatial separation in selective and divided listening. J. Acoust. Soc. Am. 119(5), 3417–3417 (2006)
13. Jack, C.E., Thurlow, W.R.: Effects of degree of visual association and angle of displacement on the "ventriloquism" effect. Percept. Motor Skill. 37, 967–979 (1973)
14. Li, Z., Duraiswami, R., Davis, L.: Recording and reproducing high order surround auditory scenes for mixed and augmented reality. In: IEEE and ACM ISMAR, pp. 240–249 (November 2004)
15. Litovsky, R.Y., Colburn, H.S., Yost, W.A., Guzman, S.J.: The precedence effect. J. Acoust. Soc. Am. 106, 1633–1654 (1999)
16. Sodnik, J., Tomazic, S., Grasset, R., Duenser, A., Billinghurst, M.: Spatial sound localization in an augmented reality environment. In: OZCHI, pp. 111–118 (2006)
17. Sundareswaran, V., Wang, K., Chen, S., Behringer, R., McGee, J., Tam, C., Zahorik, P.: 3D audio augmented reality: implementation and experiments. In: IEEE and ACM ISMAR, pp. 296–297 (October 2003)
18. Thomas, G.: Experimental study of the influence of vision on sound localization. J. Exp. Psychol. 28(2), 163–177 (1941)
19. Wallach, H., Newman, E.B., Rosenzweig, M.R.: The precedence effect in sound localization. Am. J. Psychol. 62(3), 315–336 (1949)