# The Effects of Avatar Voice and Facial Expression Intensity on Emotional Recognition and User Perception

Trinity Suma
Columbia University
New York, USA
trs2163@columbia.edu

Birate Sonia
University of Virginia
Virginia, USA
trj8ap@virginia.edu

Kwame Agyemang Baffour
Oyewole Oyekoya
City University of New York
New York, USA
kbaffour@gradcenter.cuny.edu
oyewole.oyekoya@hunter.cuny.edu

(a) Aggressor Character

(b) Bystander/Assertive Character

**Figure 1: Avatar representations with three facial expressions each: (1) low intensity; (2) medium intensity; (3) high intensity**

## ABSTRACT

The use of avatars of various rendering styles (e.g., abstract, cartoon, realistic) in virtual reality is ever-increasing. However, little is known about the effects of auditory stimuli, specifically avatar voices, on users' perceived realism. This paper aims to investigate and better understand the role of a look-alike avatar's vocal and facial expression intensity on users' perceived realism and emotional recognition using a virtual bystander scenario. Results show that avatars' vocal intensity generally affected study participants' emotional recognition while facial expression intensity affected their perceived realism. The results have implications for the perception and effectiveness of look-alike avatars in virtual environments, specifically industry training for dangerous or non-replicable situations, such as school shootings and exposure therapy.

## CCS CONCEPTS

• **Computing methodologies** → *Perception.*

## KEYWORDS

Look-alike Avatar, Bystander Intervention, Virtual Reality, Voice

**ACM Reference Format:**

Trinity Suma, Birate Sonia, Kwame Agyemang Baffour, and Oyewole Oyekoya. 2023. The Effects of Avatar Voice and Facial Expression Intensity on Emotional Recognition and User Perception. In *SIGGRAPH Asia 2023 Technical Communications (SA Technical Communications '23), December 12–15, 2023, Sydney, NSW, Australia.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3610543.3626158



*SA Technical Communications '23, December 12–15, 2023, Sydney, NSW, Australia*



ACM ISBN 979-8-4007-0314-0/23/12…$15.00
https://doi.org/10.1145/3610543.3626158

## 1 INTRODUCTION

Look-alike avatars, which are digital models of people that look like the users they represent [Frampton-Clerk and Oyekoya 2022], are preferred by users due to their human-like appearances [Pakanen et al. 2022]. A look-alike avatar's perceived realism is determined by features such as eye movement, body language, and facial gestures. Avatars have a higher perceived realism both in a static state and with full body animation and the lowest perceived realism when only their lips are synced with the audio [Frampton-Clerk and Oyekoya 2022]. However, little is known about the effects of auditory stimuli, specifically avatar voices, on users' perceived realism. Audio presence has been shown to increase users' emotional recognition [Mukashev et al. 2021], but it is unknown how varying levels of voice intensity affect this recognition. This paper aims to investigate and better understand the role of a look-alike avatar's vocal and facial expression intensity on users' perceived realism and emotional recognition using a virtual bystander scenario. Virtual reality has become increasingly present in various industries; for example, it is used to train for dangerous or impractical, non-replicable situations, such as school shootings, and psychotherapy or exposure therapy [Slater et al. 2009]. Thus, it is important to understand the effect of facial animation and voice on avatars' realism and emotional recognition to ensure that VR training is as effective as possible and that experiential fidelity is maximized.

## 2 RELATED WORK

We describe research in three areas: the role of avatars' face and body animation on users' perception in a virtual setting, the role of avatars' voices on users' perception and immersion in a virtual setting, and the use of VR in bystander scenario training and experiences. Pakanen et al. [Pakanen et al. 2022] studied how users

prefer to see other users in both AR and VR when performing a collaborative task using a telexistence system. They found that users prefer avatars that have greater photorealism and look like the user they represent. Such avatars invoked higher perceived competence and a greater sense of co-presence. Other literature suggests that an avatar's facial features and animation influence its perceived realism and the overall user experience. Frampton-Clerk and Oyekoya [Frampton-Clerk and Oyekoya 2022] found that lip animation alone, when synced to audio, incites the uncanny valley effect and, thus, a low sense of perceived realism. Contrarily, full-body and full-face animation promote the greatest perceived realism, indicating the importance of facial expressions and body movement when creating and animating lookalike avatars.

Kao et al. [Kao et al. 2021] investigated the effect of self-similar avatar voices on participants' motivation and performance in the Java programming game CodeBreakers. They found that voice similarity increased users' performance, time spent on the game, similarity identification, competence, relatedness, and immersion. However, in this study, avatars were of solid color, only designed to align with the user's gender. So, it is still unknown how voice similarity, coupled with a look-alike avatar, influences these outcomes and overall user perception in a virtual setting. In their study investigating the role of audio in emotional recognition, Mukashev et al. [Mukashev et al. 2021] found that the presence of audio increased participants' ability to identify the emotion expressed by the avatar. Nass et al. [Nass et al. 2001] studied voice emotion and news content emotion consistency in recorded and synthesized speech. They found that the emotion of a voice significantly influenced the perception of the valence of the content for both recorded and synthesized speech. Study participants also found the content more credible when the voice emotion and content emotion were not the same (e.g. a happy news story said in a sad voice).

Studies have used VR to facilitate anti-bullying training and bystander intervention training. Oyekoya et al. [Oyekoya et al. 2021] designed an interactive game prototype that allows elementary and middle school students to assume various roles (e.g. bystander, bully, victim) in a bullying scenario. Focus group feedback suggested that allowing users to have a greater role in the scenario's creation through scripting, world design, and avatar customization may increase immersion and involvement. The focus groups also suggested that adding facial expressions and animation may facilitate greater empathy. McEvoy et al. [McEvoy et al. 2016] compared users' reactions to a bullying scene through customized and non-customized VR and through video. They found that the customized VR prompted more user intervention than the non-customized VR. A follow-up study indicated that participants think they would have experienced greater empathy towards bullying victims and felt more immersed if the avatars were photorealistic.

# 3 METHODOLOGY

## 3.1 Participants

A total of 26 participants completed the anonymous survey. Two participants were under 18 years old, sixteen participants were between the ages of 18 to 24, four participants were between the ages of 25 to 34, one participant was between the ages of 45 to 54, and two participants were 55 or older. Twelve participants identified as female, thirteen identified as male, and one identified as non-binary. 11 participants had no prior knowledge of avatars or the use of avatars. 25 participants had normal or corrected-to-normal vision. Participants were recruited through online forums, social media sites including Instagram and Facebook, and direct messaging.

## 3.2 Materials

The avatars for this study were designed using Reallusion Character Creator v4.23 and the Headshot plugin v1.0. Two subjects volunteered to have their look-alike avatar representations generated. Avatars were animated and recorded using iClone v8.02 with the Motion LIVE plugin. The final videos were edited in CapCut v8.6.0.

## 3.3 Procedures and Measures

Video representations of the avatars were generated and included in an anonymous online survey distributed via Qualtrics. Both volunteers were instructed to act either aggressively or assertively at three different levels: high-intensity, medium-intensity, and low-intensity. For high-intensity expressions, the volunteers were instructed to use an overly loud vocal tone and exaggerated facial expressions. For medium-intensity expressions, the volunteers were instructed to use their normal vocal tone and facial expressions. For low-intensity expressions, the volunteers were instructed to use a quiet vocal tone and subtle facial expressions. Through written narratives and videos of avatars representing the characters in the storyline, the participants viewed a bystander intervention scenario. The participant viewed nine variations of the same avatar, each with a different combination of facial expression and voice intensity. To create each variation, we extracted the audio from the original video, thereby separating the voice intensity (audio) from the facial expressions (video). Then, all three intensities of facial expressions were mixed with the voices to create nine videos. Following each video, the participants rated the avatar's level of aggression or assertion on a 5-point scale for the aggressor and bystander characters, respectively, with (1) being the least aggressive/assertive and (5) being the most aggressive/assertive. Participants also rated their perceived realism of each avatar on a 5-point scale, with (1) being the least realistic and (5) being the most realistic. The survey aimed to gauge participants' initial perceptions of the virtual characters as soon as they finished watching the videos of the avatar representations. The survey and videos were presented in a single form, enabling the participants to re-watch videos, compare representations, and change their selections before submitting their final responses, which could not be changed after submission.

# 4 RESULTS AND ANALYSIS

The study used a repeated measures (within-subjects) design with three independent variables: (i) emotional state (aggressive and assertive); (ii) facial expression (high, medium and low intensity); and (iii) voice (high, medium and low intensity). Participants experienced all 18 (2x3x3) conditions randomly to reduce any confounding influence of the order and sequence effects such as learning or fatigue. We treat emotional state, facial expression and voice as categorical independent variables (factors), and perceived emotion (aggression/assertiveness) and perceived realism as ordinal dependent variables and conducted an ordinal logistic regression of the perceived emotion and perceived realism on the

**Table 1: Median and (mean) values for the AGGRESSIVE avatar's perceived aggression[a] and perceived realism[b] ratings. Bolded are the variations with the highest and lowest ratings.**

| | | Facial Expression | | |
|---|---|---|---|---|
| | | **High intensity** | **Medium intensity** | **Low intensity** |
| **Voice** | **High intensity** | **4 (3.78)**[a], **3 (3.04)**[b] | 3 (2.92)[a], 2 (2.31)[b] | 3 (2.88)[a], **2 (2.08)**[b] |
| | **Medium intensity** | 2 (2.38)[a], 3 (2.96)[b] | 2 (2.12)[a], 3 (2.73)[b] | 2 (1.81)[a], 2 (2.23)[b] |
| | **Low intensity** | 2 (2.04)[a], 3 (2.81)[b] | 1.5 (1.77)[a], 3 (2.96)[b] | **1 (1.15)**[a], 2 (2.42)[b] |

**Table 2: Median and (mean) values for the ASSERTIVE avatar's perceived assertion[c] and perceived realism[d] ratings. Bolded are the variations with the highest and lowest assertion ratings.**

| | | Facial Expression | | |
|---|---|---|---|---|
| | | **High intensity** | **Medium intensity** | **Low intensity** |
| **Voice** | **High intensity** | **4 (3.77)**[c], 3 (2.69)[d] | 4 (3.35)[c], 3 (2.62)[d] | 3 (3.31)[c], 2.5 (2.62)[d] |
| | **Medium intensity** | 2 (2.27)[c], 2 (2.73)[d] | 2 (1.77)[c], 2 (2.62)[d] | 2 (2.12)[c], 2 (2.5)[d] |
| | **Low intensity** | 1.5 (1.73)[c], **2 (2.08)**[d] | 1 (1.58)[c], 3 (2.69)[d] | **1 (1.50)**[c], **3 (2.78)**[d] |

independent variables (using SPSS Statistics 25 software. The proportional odds assumption of logistic regression is not violated for perceived emotion ($\chi^5 1 = 26.680, p = 0.998$) or perceived realism ($\chi^2(51) = 48.796, p = 0.562$). Pearson Chi-square test ($\chi^2(51) = 52.686, p = 0.409$; $\chi^2(51) = 46.400, p = 0.657$) and the Deviance test ($\chi^2(51) = 53.326, p = 0.385$; $\chi^2(51) = 48.796, p = 0.562$) were both non-significant for perceived emotion and perceived realism respectively, suggesting that the models exhibit good fit to the data. We treat the medium-intensity voice and medium-intensity facial expression as the reference categories for the independent variables, voice and facial expression respectively. This makes sense, as the baseline for measuring the perceived emotion/realism of the high and low voices and facial expressions, especially as the medium-intensity was the neutral way that the actors would express themselves naturally. Also, since the interactions between the voice intensity and facial expressions are important, a statistical test of model effects showed that there were significant 2-way (facial expression and voice intensity, $p < 0.05$) and 3-way interactions (emotional state, facial expression and voice intensity, $p < 0.05$). The means and medians are presented in Tables 1 and 2.

## 4.1 Emotional Recognition

Wald test shows that the independent variables (emotional state and facial expression) were not significant predictors ($p > 0.05$) of perceived emotion. However, high-intensity voice was a significant positive predictor ($p < 0.0001$) of perceived emotion. Participants were 17.686 times more likely to perceive the characters' high-intensity voice as aggressive as opposed to the medium-intensity voice. Additionally, aggression was rated highest in the aggressor character when both the facial expressions and voice were both at the highest intensity; the mean response was 3.78 and the median response was 4. 69.23% of study participants rated aggression 4 or higher for this variation of the character. Contrarily, aggression was rated lowest when the facial expressions and voice were both at the lowest intensity; the mean response was 1.15 and the median response was 1. 96.15% of participants rated aggression at 2 or lower for this variation of the character. Table 1 shows the mean and median responses for all variations of the avatar, with the highest and lowest ratings in bold. Assertion was rated highest in the bystander character when the facial expressions and voice were both at the highest intensity; the mean response was 3.77 and the median response was 4 (see Table 2). 65.38% of study participants rated assertion 4 or higher for this variation of the character. Contrarily, assertion was rated lowest when the facial expressions and voice were both at the lowest intensity; the mean response was 1.5 and the median response was 1. 92.31% of participants rated assertion at 2 or lower for this variation of the character. Table 2 shows the mean and median responses for all variations of the avatar, with the highest and lowest ratings in bold.

## 4.2 Perceived Realism

Wald test shows that the independent variables (emotional state, voice intensity and facial expression) were not significant predictors ($p > 0.05$) of perceived realism. Perceived realism was highest in the aggressive character when the facial expressions and voice were both at the highest intensity; the mean response was 3.04 and the median response was 3 (see Table 1). 26.92% of study participants rated realism 4 or higher for this variation of the character. Contrarily, perceived realism was lowest when facial expressions were at the lowest intensity but the voice was at the highest intensity; the mean response was 2.08 and the median response was 2. 73.08% of participants rated realism at 2 or lower for this variation of the character. Table 1 shows the mean and median responses for all variations of the avatar, with the highest and lowest ratings in bold. Perceived realism was highest in the assertive character when the facial expressions and voice were both at the lowest intensity; the mean response was 2.78 and the median response was 3 (see Table 2). 23.08% of study participants rated realism 4 or higher for this variation of the character. Contrarily, perceived realism was lowest when facial expressions were at the highest intensity but the voice was at the lowest intensity; the mean response was 2.08 and the median response was 2. 69.23% of participants rated realism at 2 or lower for this variation of the character. Table 2 shows the mean and median responses for all variations of the avatar, with the highest and lowest ratings in bold.

# 5 DISCUSSION

High levels of aggression and assertion were continuously associated with high facial and vocal intensity while low levels of aggression and assertion were continuously associated with low facial and vocal intensity. This is likely due to the consistency between facial and vocal intensity, allowing participants to easily recognize

the emotion presented. When the vocal and facial expression intensities were inconsistent, participants often rated their perceived levels of aggression and assertion at a 3 or lower. Furthermore, low vocal intensity frequently resulted in users ranking the aggressive avatar's aggression as lower, with at least 69.23% of participants ranking it 2 or lower for every face variation. Similarly, low vocal intensity frequently resulted in users ranking the bystander avatar's assertion as lower, with at least 84.62% of participants ranking it 2 or lower for every face variation. This may be due to the dominance of voice and auditory stimuli in users' emotional recognition. However, we did not find the same results for recognizing high levels of aggression or assertion. High vocal intensity resulted in at least 30.77% of participants ranking the avatar's aggression at a 4 or higher and 46.15% of participants ranking the bystander's assertion at a 4 or higher. This could be due to a difference in standards between participants regarding high levels of aggression and assertion but a similar standard for low levels of aggression and assertion.

Perceived realism data were more normally distributed than that collected for emotional recognition. This could be due to participants' different standards or characteristics of realism; participants could have interpreted realism as how the avatar looked or how the avatar behaved. The data suggest that facial expression intensity has a greater effect than vocal intensity on users' perceived realism, with realism decreasing as facial expression intensity lessens. This could be due to users' primarily focusing on the avatars' appearances when determining realism. For the aggressive character, realism was highest when both the facial expressions and voice were at the highest intensity. This may be because aggression is an intense emotion, therefore, requiring intense facial expressions and vocals to appear realistic. Realism was lowest when a high-intensity voice was coupled with a low-intensity facial expression. This mismatch could have incited the uncanny valley effect, resulting in a lower sense of perceived realism, as found by Frampton-Clerk and Oyekoya [Frampton-Clerk and Oyekoya 2022]. For the assertive character, realism was highest when both the facial expressions and voice were at the lowest intensity. This result was the opposite of what we found for aggressive avatar and could have been because the participants thought the high-intensity variations were overly forced and, thus, unrealistic, but further research should investigate this preference. Realism was lowest when the low-intensity voice was coupled with a high-intensity facial expression. This mismatch could have also incited the uncanny valley effect and, thus, a low sense of perceived realism.

## 6 CONCLUSION

In this paper, we conducted a study to investigate how look-alike avatars' facial expressions and vocal intensity affect users' emotional recognition and perceived realism. The results show vocal intensity having a greater effect on emotional recognition and facial expressions having a greater effect on perceived realism. This suggests that facial and vocal intensity may be crucial features in maximizing experiential fidelity in virtual reality.

Several limitations should be considered when generalizing the results of this study. First was the creation of a male avatar as the aggressor character and a female avatar as the bystander character. Since men are more often perpetrators of violence, including relationship and sexual violence, males are often depicted as aggressors while females are often victims, even though men are more consistently found to be at risk of violent victimization [Krug et al. 2002]. This prevalence may induce unconscious bias in study participants, causing the aggressor's emotions to appear more intense. Future work may investigate how emotional recognition and perceived realism vary using a female aggressor and a male bystander or avatars of different races. Additionally, this study is limited by the volunteers' interpretations of aggression and assertion. Future work may examine how other volunteers' interpretations of high, medium, and low-intensity aggression and assertion influence participants' emotional recognition and perceived realism. Furthermore, another limitation was the sample size of twenty-six participants, given that it was an online survey. Participant feedback indicated that the survey crashed at the end for some users, which could have been due to its desktop optimization rather than mobile. Many incomplete responses were removed, thus, decreasing the sample size. Of the final sample size, more than half identified with the 18-24 age range. Future work may aim to gauge a large sample size with a more diverse age range to more accurately gauge all potential users' perceptions. Future work may also examine if the demographic information we collected on gender, age, and avatar familiarity impacted participants' responses, as this requires a much larger sample. As this paper only focused on the role of facial expressions and voice on emotional recognition, future work may investigate the role of full-body animation on emotional recognition, since previous work has shown that it has a positive effect on perceived realism [Frampton-Clerk and Oyekoya 2022].

## ACKNOWLEDGMENTS

This work was supported by the National Science Foundation, Research Experience for Undergraduates Award No. 2050532.

## REFERENCES

Aisha Frampton-Clerk and Oyewole Oyekoya. 2022. Investigating the Perceived Realism of the Other User's Look-Alike Avatars. In *Proceedings of the 28th ACM Symposium on Virtual Reality Software and Technology*. 1–5.

Dominic Kao, Rabindra Ratan, Christos Mousas, and Alejandra J Magana. 2021. The effects of a self-similar avatar voice in educational games. *Proceedings of the ACM on Human-Computer Interaction* 5, CHI PLAY (2021), 1–28.

Etienne G Krug, James A Mercy, Linda L Dahlberg, and Anthony B Zwi. 2002. The world report on violence and health. *The lancet* 360, 9339 (2002), 1083–1088.

Kelly A McEvoy, Oyewole Oyekoya, Adrienne Holz Ivory, and James D Ivory. 2016. Through the eyes of a bystander: The promise and challenges of VR as a bullying prevention tool. In *2016 IEEE Virtual Reality (VR)*. IEEE, 229–230.

Dinmukhamed Mukashev, Merey Kairgaliyev, Ulugbek Alibekov, Nurziya Oralbayeva, and Anara Sandygulova. 2021. Facial expression generation of 3D avatar based on semantic analysis. In *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. IEEE, 89–94.

Clifford Nass, Ulla Foehr, Scott Brave, and Michael Somoza. 2001. The effects of emotion of voice in synthesized and recorded speech. In *Proceedings of the AAAI symposium emotional and intelligent II: The tangled knot of social cognition*. AAAI North Falmouth, MA.

Oyewole Oyekoya, Jan Urbanski, Yaroslava Shynkar, Arifa Baksh, and Margaret Etsaghara. 2021. Exploring First-Person Perspectives in Designing a Role-Playing VR Simulation for Bullying Prevention: A Focus Group Study. *Frontiers in Virtual Reality* 2 (2021), 672003.

Minna Pakanen, Paula Alavesa, Niels Van Berkel, Timo Koskela, and Timo Ojala. 2022. "Nice to see you virtually": Thoughtful design and evaluation of virtual avatar of the other user in ar and vr based telexistence systems. *Entertainment Computing* 40 (2022), 100457.

Mel Slater, Pankaj Khanna, Jesper Mortensen, and Insu Yu. 2009. Visual realism enhances realistic response in an immersive virtual environment. *IEEE computer graphics and applications* 29, 3 (2009), 76–84.