Effects of Body Type and Voice Pitch on Perceived Audio-Visual Correspondence and Believability of Virtual Characters

Luchcha Lam
lam124@purdue.edu
Purdue University
West Lafayette, Indiana, USA

Minsoo Choi
choi714@purdue.edu
Purdue University
West Lafayette, Indiana, USA

Magzhan Mukanova
mmukanov@purdue.edu
Purdue University
West Lafayette, Indiana, USA

Klay Hauser
hauser12@purdue.edu
Purdue University
West Lafayette, Indiana, USA

Fangzheng Zhao
fangzheng.zhao@psych.ucsb.edu
University of California, Santa Barbara
Santa Barbara, California, USA

Richard Mayer
mayer@ucsb.edu
University of California, Santa Barbara
Santa Barbara, California, USA

Christos Mousas
cmousas@purdue.edu
Purdue University
West Lafayette, Indiana, USA

Nicoletta Adamo
nadamovi@purdue.edu
Purdue University
West Lafayette, Indiana, USA

Figure 1: Female (left) and male (right) body type variations. From left to right for both female and male virtual characters: ectomorph, mesomorph, and endomorph.

ABSTRACT

We examined the effects of virtual characters' body type and voice pitch on perceived audio-visual correspondence and believability. For our within-group study ($N = 72$), we developed nine experimental conditions using a 3 (body type: ectomorph vs. mesomorph vs. endomorph body types) × 3 (voice pitch: low vs. medium vs. high fundamental frequency [F0]) design. We found statistically significant main effects from voice pitch and statistically significant interaction effects between a virtual character's body type and voice pitch on both the level of perceived audio-visual correspondence and believability of female and male virtual characters. For female virtual characters, we also observed an additional statistically significant main effect from body type and a statistically significant interaction effect between the participant's biological sex and the virtual character's voice pitch on both perceived audio-visual correspondence and believability. Moreover, the results show that perceived believability is highly correlated to perceived audio-visual correspondence. Our findings have important practical implications in applications where the virtual character is meant to be an emotional or informational guide that requires some level of perceived believability, as the findings suggest that it is possible to enhance the perceived believability of the virtual characters by generating appropriate voices through pitch manipulation of existing voices.

SAP '23, August 05–06, 2023, Los Angeles, CA, USA

ACM ISBN 979-8-4007-0252-5/23/08.
https://doi.org/10.1145/3605495.3605791

CCS CONCEPTS

• **Computing methodologies** → **Animation**; **Perception**.

KEYWORDS

virtual characters, voice pitch, facial features, body dimensions, perceived audio-visual correspondence, perceived believability

ACM Reference Format:

Luchcha Lam, Minsoo Choi, Magzhan Mukanova, Klay Hauser, Fangzheng Zhao, Richard Mayer, Christos Mousas, and Nicoletta Adamo. 2023. Effects of Body Type and Voice Pitch on Perceived Audio-Visual Correspondence and Believability of Virtual Characters. In *ACM Symposium on Applied Perception 2023 (SAP '23), August 05–06, 2023, Los Angeles, CA, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3605495.3605791

1 INTRODUCTION

In virtual characters, audio-visual components play a crucial role in creating a realistic and engaging user experience. Audio plays an important role in human interactions in the form of voice, which is the most common and expedient way of communication [Wani et al. 2021]. Moreover, voice conveys personal information such as social status, personality traits, and the speaker's emotional state [Kreiman and Sidtis 2011; Zhang 2016]. Much like how we look, the way we sound comprises overlapping data about who we are to the extent that people can imagine a speaker's face from their voice and vice versa [Kao et al. 2022; Mavica and Barenholtz 2013]. Finding the right voice is crucial to the level that a mismatch in the realism of virtual human face and voice creates the feeling of discomfort [Mitchell et al. 2011] or uncanniness [Tinwell et al. 2010].

Recent neuroscience research has shown a link between auditory and visual cues in human perception [Belin et al. 2004] and that the cues are processed interchangeably in the cortex [Young et al. 2020]. Studies on humans have shown correlations between facial features and the acoustic features of the voice [Lu et al. 2021; Mavica and Barenholtz 2013; Oh et al. 2019] as well as correlations between the body and the voice [Evans et al. 2006; Pawelec et al. 2020; Pisanski et al. 2016]. Additionally, it is known that age and gender [Yamauchi et al. 2022], and ethnicity [Rakić et al. 2011] affect how a person sounds and speaks. Yet, it is inconclusive whether a person's physical dimensions can affect their voice, making the role of physical dimensions on voice unknown. Thus, we leveraged these audio-visual correspondences in humans, which are reflected in our stimuli, to study the impact of perceived audio-visual correspondence and believability of virtual characters. The perceived audio-visual correspondence refers to the match between the virtual character's voice pitch, expressed as the fundamental frequency (`F0`), and the virtual character's face and body dimensions measured by the facial width, jaw length, neck girth, chest girth, and waist and hip girth. We define perceived believability as the likeliness that the virtual character produces the given voice.

Based on the previous relationships between the acoustic features of the human voice and the physical characteristics of the speaker, we conducted a 3 (body type: ectomorph vs. mesomorph vs. endomorph body types; see Figure 1) × 3 (voice pitch: low vs. medium vs. high fundamental frequency [`F0`]; see Table 2 in the supplementary material document) within-group study to explore the significance of this relationship on virtual characters. This study aimed to answer the following research questions:

- **RQ1.1:** To what extent does the body type affect the perceived audio-visual correspondence for female and male virtual characters?
- **RQ1.2:** To what extent does the body type affect the perceived believability of female and male virtual characters?
- **RQ2.1:** To what extent does the voice pitch affect the perceived audio-visual correspondence for female and male virtual characters?
- **RQ2.2:** To what extent does the voice pitch affect the perceived believability of female and male virtual characters?
- **RQ3.1:** To what extent does the body type × voice pitch interaction impact the perceived audio-visual correspondence for female and male virtual characters?
- **RQ3.2:** To what extent does the body type × voice pitch interaction impact the perceived believability for female and male virtual characters?
- **RQ4:** Is there a correlation between perceived audio-visual correspondence and believability?
- **RQ5:** To what extent does the perceived audio-visual correspondence and believability of body types and voice pitch depend on the participant's sex?

2 RELATED WORK

Believable Virtual Characters. Researchers have been working on creating believable virtual characters in various aspects such as graphics [Walshe et al. 2003], interactions [Riedl and Stern 2006], and emotional abilities [Marsella and Gratch 2003]. A believable autonomous agent is a life-like system with reactivity and interactivity that is able to make appropriate decisions [Riedl and Stern 2006]. Loyall et al. [Loyall et al. 1997] identified key features of a believable virtual agent, including a unique personality, emotional awareness, self-motivation, responsiveness, consistency of expression, and the illusion of life. Virtual agents that demonstrate environmental awareness and interaction awareness were perceived as more believable as they were able to react and exist in the correct social context [Bogdanovych et al. 2016]. Correspondingly, Doyle [Doyle 2002] differentiated realism and believability in virtual characters by stating that a believable virtual character is not necessarily a real character but is real in the context of its environment. Thomas and Johnston [Thomas and Johnston 1981] claimed that personality match, even when the virtual character does not look realistic, can make drawings real in animation. An example of this is when audiences relate to stylized hand-drawn characters with consistent personalities as they appear to be real to the audience [Han 2009]. Hence, published works agreed upon using the extent to which a viewer engages and empathizes with a virtual character as a measurement of perceived believability [Han 2009; Niewiadomski and Pelachaud 2011].

Perception of Virtual Character's Voices. Virtual characters can have real actors' voices, synthesized voices, or a mixture of both. Studies on text-to-speech and copy-synthesis methods reported that synthesized voices are perceived as less sympathetic [Thomas et al. [n. d.]] and less preferable [Cabral et al. 2017] when compared to real human voices. Despite the negative feedback on certain

synthesis methods, other studies revealed that the perceived naturalness in voice is related to the speaker's distinct characteristic rather than the realism of the voice. Unnatural-sounding voices do not affect a virtual character's social presence nor empathetic responses from the audiences [Aylett et al. 2017; Higgins et al. 2022]. Yet, a mismatch in realism between the voice and the virtual character's appearance can develop a sense of uncanniness, resulting in discomfort for the viewer [Higgins et al. 2022; Mitchell et al. 2011]. Ferstl et al. [Ferstl et al. 2021b] study with animated virtual agents highlighted the importance of believable voices in virtual characters by illustrating that realism in voice is preferable over the realism of appearance when they produce perceptual mismatches. Works in this area [Ferstl et al. 2021b; Kao et al. 2022] demonstrated that matching the appropriate voices for virtual characters is crucial because vocal characteristics and their appropriateness can influence human perception of a virtual character.

The Role of Voice in Human Recognition. Voice contains information about the speaker and influences our perceptions of the speaker. With virtual characters, viewers can perceive agreeableness and emotional stability through their speech [Thomas et al. [n. d.]]. The human voice could be a predictor for the speaker's attributes such as identity [Maguinness et al. 2018], age [Grzybowska and Kacprzak 2016], gender [Junger et al. 2013], size [Smith et al. 2005], emotion [Zhang et al. 2019], personality [Aronovitch 1976], weight [de Souza and dos Santos 2018], and height [Pisanski et al. 2014]. Furthermore, voice influences how we perceive the speaker, such as attractiveness [Collins and Missing 2003; Pisanski et al. 2016], femininity and masculinity [Cartei and Reby 2013; Coleman 1976], social status [Zhang 2016], and ethnicity [Oh et al. 2019]. All in all, voice plays a critical role in the embodiment of a human, as seen with actors, who have been engaging their voices with their faces and bodies to build a holistic portrayal of a virtual character [Berry et al. 2022].

Humans unconsciously associate faces with voices in identity recognition [Joassin et al. 2011]. Yet, there is no strong, consistent correlation between the face and the voice. While some studies reported that people could only match unfamiliar voices to dynamically articulating faces and not to static photographs [Kamachi et al. 2003; Lachs and Pisoni 2004], some also showed that people could match unfamiliar voices to faces with slightly greater than chance accuracy, with better performance for dynamic faces than for static faces [Mavica and Barenholtz 2013; Smith et al. 2016]. Discrepancies between the results in studies related to face and voice correlation may be due to the lack of diversity [Kamachi et al. 2003] in their pool of participants and the inconsistencies in their methodologies such as sequential [Lachs and Pisoni 2004] versus simultaneous presentation of stimuli [Mavica and Barenholtz 2013]. Presentation order for the stimuli may have caused temporal biases since research suggested that memory for dynamic facial images is better than that of static images [Christie and Bruce 1998]. Additionally, in real interactions, faces and voices are also commonly perceived simultaneously [Smith et al. 2016].

Attributes of the voice are related to the physical facial and body structure. People with longer and wider faces are often associated with lower voices [Macari et al. 2014, 2017], but some studies only reported this correlation for certain vowels [Bommarito 2019; Macari et al. 2015]. The agreed-upon trends in this area suggest that people with larger neck girth [Pawelec et al. 2020], higher body mass [Cartei and Reby 2013], and older age are associated with lower voices [Cartei and Reby 2013; Eichhorn et al. 2018; Hatano et al. 2012]. Literature also suggests that hip and chest circumferences, and waist circumference negatively and positively correlated to the fundamental frequency (`F0`), respectively [Evans et al. 2006; Hughes et al. [n. d.]; Pawelec et al. 2020; Pisanski et al. 2016]. The waist-to-hip ratio also corresponds to the hormone levels [Hughes et al. [n. d.]; Mondragón-Ceballos et al. 2015] that influence the vocal tract and voice qualities [Pawelec et al. 2020].

Although the length of the vocal tract, responsible for speech production, seems to correlate with the speaker's height [Barsties et al. 2016; Fitch and Giedd 1999; Pisanski et al. 2014, 2016] in which a taller body is associated with a longer vocal tract and lower fundamental frequencies [Fitch and Giedd 1999; Pisanski et al. 2014], inconsistent results still exist for voice correlation to body height, especially in different sexes [Barsties et al. 2016; Hatano et al. 2012]. There are several external and biophysical influential factors on voice, especially on the undisclosed fundamental frequency, such as age [Fitch and Giedd 1999], time of day [Garrett and Healey 1987], speaking context [Zraick et al. 2006], languages [Altenberg and Ferrand 2006], emotions [Breitenstein et al. 2001], hormonal changes [D'haeseleer et al. 2012], and even habits such as smoking [Guimarães and Abberton 2005]. As a result, voice condition varies significantly. Research in this area is still limited by the small number of voice conditions [Zhang 2016] and the relatively small number of variables [Berry et al. 2022] explored in each study. Therefore, these voice parameters could explain only a small amount of the variation in human body measurements [Pisanski et al. 2016].

3 MATERIALS AND METHODS

3.1 Participants

We conducted two *a priori* power analyses on the collected perceived audio-visual correspondence and believability ratings from 10 participants using R-Studio software. Based on a .61 effect size for the perceived audio-visual correspondence rating and .43 effect size for the perceived believability rating and an $\alpha = .05$, to achieve an 80% power, the analyses recommended a minimum of 70 participants. We recruited 72 participants from our university campus for our within-group study. Our pool of participants consisted of 25 females and 47 males with ages ranging from 18 to 40 ($M = 24.42$, $SD = 4.01$). Our participants self-reported their English fluency to be at least at working proficiency and have no ear or eye-related disabilities.

3.2 Virtual Character Generation and Animation

We used Reallusion's Character Creator 4[1] with its default content library and the *Working Class Heroes*[2] collection downloaded from Reallusion's content store to create our virtual characters. We designed six virtual characters (see Figure 1). Each virtual character had three variations of different body dimensions based on

[1]www.reallusion.com/character-creator
[2]www.reallusion.com/ContentStore/Character-Creator/Pack/Working-Class-Heroes

the *Heath-Carter Somatotypes* [Carter and Heath 1990]: ectomorph, mesomorph, and endomorph (see Table 1 in the supplementary material document). We used the somatotypes as the base guidelines for our virtual character as we tried to capture the major human body types. We limited our study to American English-speaking virtual characters of Caucasian descent to match the source of our base voice corpus. We selected brown hair for our virtual characters because it is the most common hair color in America [hai [n. d.]]. We kept the height of the virtual characters constant as height's contribution to voice is inconsistent and difficult to distinguish through a computer screen. Age is also constant within the range for the most stable voice (30-50 years old) [Abitbol et al. 1999].

We wrote Python scripts in Autodesk Maya 2022[3] to evaluate that the virtual character's mesh dimensions match the body measurements presented in human research. There is a fixed difference in the mesh measurement between the body variations to ensure equal influences. The mesh measurements include the facial width, the jaw length, and the neck, chest, waist, and hip girth. The facial width is calculated by intersecting a plane with the head at the level of the bilateral zygion points [Macari et al. 2017] and calculating the Euclidean distance between the two furthest points from that intersection. The jaw length is estimated with the Euclidean distance between the vertices representing the condylion and the lowest point of the mandibular symphysis from the side view [Macari et al. 2014]. The neck, chest, waist, and hip girth were estimated by intersecting those parts of the model with planes as shown in Figure 2 and calculating the circumference of the convex hull formed by the polygonal chain of the intersection points [Wuhrer and Shu 2013].

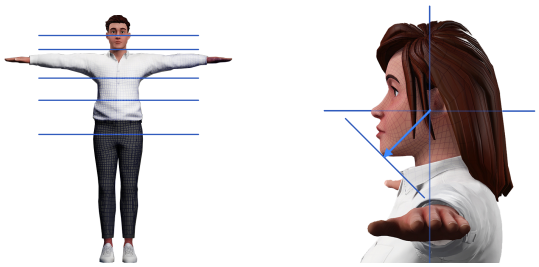

Figure 2: Left: The result of our custom Python script to create intersecting planes used to measure the different mesh dimensions: facial width, neck girth, chest girth, and waist and hip girth in this order from top to bottom. Right: The result of our custom Python script to measure the jaw length.

We kept the other components of our animated virtual characters, such as the outfit [Lightstone et al. 2011], animation [Badathala et al. 2018], view [Baranowski and Hecht 2018], and lighting [Wisessing et al. 2020] constant to limit external influences on the participant's perception. We placed our virtual characters, dressed in neutral black and white, in the `Default_CC4` atmospheric lighting. Our virtual characters have the same textures, skin color, and hair colors and were presented on a 50% grey background for equal contrast (see Figure 3). The virtual characters were animated in Reallusion's iClone 8,[4] which allowed us to generate custom animations that can be applied to multiple virtual characters, guaranteeing that all of our virtual characters were animated in the same way. In addition to the lip-sync, we applied extra animations at low intensity, including facial movements to complement the lip-sync, idle movements, and two-eye blinks to avoid uncanny movements.

Figure 3: Full-body (left) and closeup (right) stimuli sample in 16:9 aspect ratio.

3.3 Voice Pitch Manipulation

The audio stimuli were manipulated from the UW/NU corpus [uwc 2017] of phonetically balanced sentences for high-quality audio. We selected Pacific Northwestern male and female recordings with minimal prosody and a neutral comprehensible pace. The audio file we selected said *"two blue fish swam in a tank."* We approached voice as how the virtual character sounds, an instrument of expression, rather than speech, which is a conveyor of meaning [Edwards and Newell 2005]. Therefore, our study excludes how the virtual character speaks, such as prosody and accent. Manipulating a single base audio for each sex instead of using human voice actors minimized the differences among the other vocal attributes and ensured the same speech content across the `F0` variations. Moreover, manipulating an existing audio file instead of using human voice actors also reflects a practical alternative to generating variations of voices when resources, such as voice actors, are limited. The `F0` values (see Table 2 in the supplementary material document) of the two selected base voices were then manipulated to three discrete values to produce the different voice variations (low, medium, and high). The minimum `F0`s, the maximum `F0`s [Baken and Orlikoff 2000], and the average between the two `F0`s for each sex was applied to the low, high, and medium voice variation, respectively. The medium pitch level also corresponds to the mean `F0` from Pisanski's subjects [Pisanski et al. 2016].

We considered fundamental frequency (`F0`) over formant frequency because it consists of a single number that is simpler for manipulation and is shown to be the most important feature used in voice perception [Gelfer and Mikos 2005]. The `F0` is adjusted using Praat's Vocal Toolkit's functions `Change Pitch Median` and `Variation`.[5] We validated the manipulated fundamental frequency with Praat's `Voice Report` function where the pitch contour settings were set to 75Hz and 300Hz for the floor and ceiling of male voices, and 100Hz and 500Hz for female voices. Setting the floor and ceiling of the pitch range is a recommended technical requirement to increase sampling resolution for the most accurate extraction in pitch analysis [pra 2019].

[3] www.autodesk.com/products/maya
[4] www.reallusion.com/iclone
[5] https://www.praatvocaltoolkit.com/change-pitch.html

After manipulating the F0, we normalized the audio files with Audacity's Loudness Normalization function (LUF)[6] to the standard's recommendation for loudness normalization issued by the European Broadcasting Union of 23 LUF [EBU-Recommendation 2011]. Likewise, the audio files used in the calibration process were also normalized to 23 LUF. Finally, we assume that the reverberation of our audio file does not affect the participants' perception of the virtual character's audio-visual coherence. The virtual characters were placed in front of a solid grey background; therefore, the environment is ambiguous and no reverberation is adjusted.

3.4 Study Details

For our study, we selected attributes that were shown to significantly affect the perceived characteristic of the human voice, which includes the facial width, jaw length, neck girth, chest girth, waist girth and hip girth, and the acoustic parameter: fundamental frequency (F0) [Macari et al. 2017; Pisanski et al. 2016]. To assess the correlation between the virtual characters' body dimensions and the pitch of the virtual characters' voices, and the correlation's affect on the perceived believability, we created two virtual characters of each biological sex (male and female), each with three different body type variations that are representative of the body measurements as shown in Table 1 in the supplementary material document, and three voice variations with different pitch levels as shown in Table 2 in the supplemental material document. Following the full-factorial design, we combined all levels of each factor to create 18 animated videos (nine with male and nine with female virtual characters) with audio, each representing a voice-body combination.

3.5 Measurements

We created a Qualtrics survey with our pre-survey calibration, stimuli, survey questions, and post-survey demographic questions. The pre-survey questions ensured that the equipment worked and that our participants had the required level of English proficiency and no eye- or ear-related disabilities. Our study consisted of eight rating questions and three optional free-response questions, giving us both quantitative and qualitative data. Questions evaluating the perceived audio-visual correspondence focused on the "match" between the voice pitch and the virtual character's physical appearance. Meanwhile, questions evaluating the perceived believability were based on Loyall et al.'s [Loyall et al. 1997] definition of believability, and focused on the expectations and the uniqueness of the match. Some questions were designed to assess similar values with different wordings to alleviate the possible discrepancies in interpretations. The questions were organized into head-related and body-related questions. The virtual characters were simultaneously presented with a close-up and a full-body view (see Figure 3) of the animation. Tables 1-3 in Appendix A show the questions grouped by the dependent variable they collected along with the animation view that was shown alongside them in the survey. Each animation had a frontal view alongside a 3/4 view to show dimensionality.

[6]https://manual.audacityteam.org/man/loudness_normalization.html

3.6 Procedure

We sent out a recruitment email in which volunteers can schedule a time to participate in our study. The study was conducted in person in our department's laboratory with a DELL P2722H monitor and a Sennheiser's on-ear HD25 DJ headphones. Upon arrival, the research team informed the participants of the instructions for the study and presented them with the consent form approved by the university's Institutional Review Board. First, we asked the participants to select whether they had eye- or ear-related disorders that may interfere with their ability to perceive the stimuli. If they selected yes, we would end the study. After that, the participants were prompted to write down what they heard from the two calibration audio files, one of a female speaker and one of a male speaker, saying, *"rice is often served in round bowls."* Next, the participants proceeded to watch all 18 videos of the treatment combinations in randomized order and rated the level of the perceived correspondence between the virtual characters' voice and body type, and the voice-body pair's perceived believability on a 5-point Likert scale (1: low - 5: high perceived audio-visual correspondence or believability). The participants could not return to previous survey pages after proceeding to another. After the study, the participants filled in their demographic information in the survey form. The study lasted no more than 50 minutes.

4 RESULTS

We performed our statistical analyses using R-Studio software. The normality assumptions were validated with QQ plots of the residuals. We used the Restricted maximum likelihood (REML) for fitting estimation and Satherthwaite's method from the ImerTest package [Kuznetsova et al. 2017] to calculate the significance. For simplicity of model interpretation, we only included two-factor interaction terms (sex × body type, sex × voice pitch, and voice pitch × body type), where sex was the participant's biological sex. We categorized our rating questions into two groups for each of our dependent variables, perceived audio-visual correspondence and believability. We ran two Linear Mixed Effect (LME) models on the data collected for our dependent variables to assess whether there are any statistically significant main and interaction effects between the virtual character's body type, voice pitch level, and the participant's biological sex on the dependent variables.

Since we were interested in the treatment effects of the virtual character's body type, voice, and participant's sex on the whole population, we treated participants as being randomly selected from the population whose characteristics we would like to estimate. For both analyses, we included the participant's biological sex, the virtual character's body type, and the virtual character's voice pitch as the fixed effects and participants as a random effect. We provide a detailed explanation of the LME model in the supplementary material document. Finally, we ran post hoc multiple pairwise comparisons using t-tests to assess each treatment-level combination. We provide descriptive statistics in Table 3-6 in the supplementary material document.

4.1 Female Virtual Characters

The results for female virtual characters showed similar trends in the level of perceived audio-visual correspondence (AVC) and believability (B), where we observed statistically significant main effects from body type (AVC: $F[2, 564] = 12.70$, $p < .001$, B: $F[2, 564] = 9.56$, $p = .001$) and voice pitch (AVC: $F[2, 564] = 38.37$, $p < .001$, B: $F[2, 564] = 45.63$, $p < .001$), as well as statistically significant interaction effects between body type and voice pitch (AVC: $F[4, 564] = 3.88$, $p = .004$, B: $F[4, 564] = 4.95$, $p = .001$), and between participant's sex and voice pitch (AVC: $F[2, 564] = 5.90$, $p = .003$, B: $F[2, 564] = 4.68$, $p = .010$).

4.1.1 Body Type. Our female participants rated the female ectomorph with a high-pitched voice to have statistically significant lower perceived audio-visual correspondence ratings and were less believable than when paired with a low-pitched voice (AVC: $t[564] = -3.00$, $p = .004$, $B : t[564] = -2.92$, $p = .005$) and a medium-pitched voice (AVC: $t[564] = -4.12$, $p < .001$, $B : t[564] = -4.80$, $p < .001$). Correspondingly, our male participants also rated the ectomorphic female virtual character with a high-pitched voice to be significantly less believable than that with a medium-pitched voice ($t[564] = -2.52$, $p = .016$). While only male participants rated the female ectomorph with a low-pitched voice to have marginally lower perceived audio-visual correspondence than the ectomorph with a medium-pitched voice ($t[564] = -2.03$, $p = .053$).

Both male and female participants gave statistically significant perceived audio-visual correspondence and believability ratings for female mesomorphs when paired with voices of different pitch levels. Both agreed that the female mesomorph with a high-pitched voice had lower perceived audio-visual correspondence and was significantly less believable than when paired with a low-pitched voice (AVC: male participants: $t[564] = -3.92$, $p < .001$; female participants: $t[564] = -6.40$, $p < .001$. B: male participants: $t[564] = -4.43$, $p < .001$; female participants: $t[564] = -6.13$, $p < .001$). Additionally, both male and female participants rated mesomorphs with a high-pitched voice to have statistically significant lower audio-visual correspondence and were significantly less believable than when paired with a medium-pitched voice (AVC: male participants: $t[564] = -3.13$, $p = .003$; female participants: $t[564] = -5.18$, $p < .001$. B: male participants: $t[564] = -3.71$, $p < .001$; female participants: $t[564] = -5.81$, $p < .001$). However, there were no statistically significant contrasts in the perceived believability ratings for female mesomorphs when paired with medium- or low-pitched voices.

Both male and female participants gave statistically significant ratings for the endomorphic female virtual character when paired with voices of different pitch levels. The female endomorph was perceived as significantly more believable with significantly higher audio-visual correspondence when paired with a low-pitched voice than with a high-pitched voice (AVC: male participants: $t[564] = -4.15$, $p < .001$; female participants: $t[64] = -6.60$, $p < .001$. B: male participants: $t[564] = -5.25$, $p < .001$; female participants: $t[564] = -6.83$, $p < .001$). Similarly, the perceived audio-visual correspondence and believability ratings were also significantly lower for the female endomorph with a high-pitched voice than for that with a medium-pitched voice (AVC: male participants: $t[564] = -2.39$, $p = .023$; female participants: $t[564] = -4.54$, $p < .001$. B: male participants: $t[564] = -2.23$, $p = .034$; female participants: $t[564] = -4.54$, $p < .001$). Only female participants rated the female endomorph to have a marginally higher correspondence when paired with a low-pitched voice than when paired with a medium-pitched voice ($t[564] = 2.05$, $p = .050$), female endomorph with a low-pitched voice had statistically significant higher perceived believability scores than the female endomorph with a medium-pitched voice (male participants: $t[564] = 3.03$, $p = .004$; female participants: $t[564] = 2.23$, $p = .029$).

4.1.2 Voice Pitch. We observed no statistically significant differences in perceived audio-visual correspondence and believability ratings for female virtual characters with a low-pitched voice when paired with the different body types. For female virtual characters with a high-pitched voice, both male and female participants gave statistically significant higher perceived audio-visual correspondence and believability ratings for ectomorphs than for endomorphs (AVC: male participants: $t[564] = 4.68$, $p < .001$; female participants: $t[564] = 3.31$, $p = .001$. B: male participants: $t[564] = 4.17$, $p < .001$; female participants: $t[564] = 2.98$, $p = .005$) and mesomorphs (AVC: male participants: $t[564] = 4.00$, $p < .001$; female participants: $t[564] = 3.31$, $p = .006$. B: male participants: $t[564] = 3.85$, $p < .001$; female participants: $t[564] = 2.51$, $p = .016$). However, the perceived audio-visual correspondence and believability ratings for female virtual characters with high-pitched voice does not significantly change when paired with endomorphic or mesomorphic body types.

The female virtual character with a medium-pitched voice was perceived to have statistically significant higher perceived audio-visual correspondence and believability from both male and female participants when paired with an ectomorphic body than with an endomorphic body (AVC: male participants: $t[564] = 4.25$, $p < .001$; female participants: $t[564] = 2.95$, $p = .005$. B: male participants: $t[564] = 4.47$, $p < .001$; female participants: $t[564] = 3.15$, $p = .003$). In addition, only male participants rated virtual characters with a medium-pitched voice to have statistically significant higher perceived audio-visual correspondence and believability when paired with ectomorphs than when paired with mesomorphs (AVC: $t[564] = 2.82$, $p = .007$. B: $t[564] = 2.66$, $p = .011$).

4.1.3 Participant's Biological Sex. Ectomorphic female virtual characters with a high-pitched voice were significantly less believable with significantly less audio-visual correspondence for our female participants than for our male participants (AVC: $t[216] = -2.00$, $p = .023$. B: $t[263] = -2.86$, $p = .002$). Additionally, the perceived believability ratings for the female endomorph with a high-pitched voice were also significantly lower for our female participants than for our male participants ($t[263] = -2.26$, $p = .013$). Mesomorphic female virtual characters with high-pitched voices were also significantly less believable for our female participants than for our male participants ($t[263] = -2.16$, $p = .016$).

4.2 Male Virtual Characters

Our results showed a statistically significant main effect from voice pitch (AVC: $F[2, 565] = 14.92$, $p < .001$. B: $F[2, 565] = 16.99$, $p < .001$) and a statistically significant interaction effect between

body type and voice pitch (AVC: $F[4, 565] = 5.20$, $p < .001$. B: $F[4, 565] = 4.86$, $p < .001$) on both the level of perceived audio-visual correspondence and believability.

4.2.1 Body Type. Both male and female participants gave the male ectomorph significantly higher perceived audio-visual correspondence and believability ratings when paired with a medium-pitched voice than with a low-pitched voice (AVC: male participants: $t[565] = -3.74$, $p < .001$; female participants: $t[565] = -2.88$, $p = .006$. B: male participants: $t[565] = -4.09$, $p < .001$; female participants: $t[565] = -2.81$, $p = .006$). Female participants also gave significantly lower perceived audio-visual correspondence and believability ratings for the male ectomorph when paired with a high-pitched voice than when paired with a medium-pitched voice (AVC: $t[565] = -2.31$, $p = .027$. B: $t[565] = -2.17$, $p = .038$). Meanwhile, only male participants gave statistically significant higher perceived audio-visual correspondence and believability scores for ectomorphs with a high-pitched voice than for the one with a low-pitched voice (AVC: $t[565] = 2.53$, $p = .016$. B: $t[565] = 2.71$, $p = .009$).

Male mesomorphs got statistically significant lower perceived audio-visual correspondence and believability ratings from male and female participants when paired with high-pitched voice than when paired with other pitch levels (AVC: male participants high-low: $t[565] = -2.62, p = .012$; female participants high-low: $t[565] = -3.81$, $p < .001$; male participants high-medium: $t[565] = -4.48$, $p < .001$; and female participants high-medium: $t[565] = -5.09$, $p < .001$. B: male high-low: $t[565] = -2.37$, $p = .024$; female high-low: $t[565] = -3.68$, $p < .001$; male high-medium: $t[565] = -4.66$, $p < .001$; female high-medium: $t[565] = -4.96$, $p < .001$). However, there were no statistically significant differences between perceived audio-visual correspondence ratings for the male mesomorphs with either low- or medium-pitched voices. Only the perceived believability ratings from male participants were significantly lower for the mesomorph with a low-pitched voice than for the mesomorph with a medium-pitched voice ($t[565] = -2.30$, $p = .029$).

Only female participants gave statistically significant lower perceived audio-visual correspondence and believability ratings for the endomorphic male virtual character when paired with the high-pitched voice than when paired with the medium-pitched voice (AVC: $t[565] = -2.26$, $p = .034$. B: $t[565] = -2.56$, $p = .014$) or low-pitched voice (AVC: $t[565] = -2.22$, $p = .031$. B: $t[565] = -2.40$, $p = .022$). However, unlike the perceived audio-visual correspondence rating for the female endomorph, there were no statistically significant differences between the perceived audio-visual correspondence nor the perceived believability for the endomorphic male virtual character when paired with low- or medium-pitched voice.

4.2.2 Voice Pitch. Our male participants perceived the male virtual character with a high-pitched voice and an ectomorphic body to have significantly more audio-visual correspondence and was significantly more believable than a male virtual character with a high-pitched voice and a mesomorphic body (AVC: $t[565] = 2.54$, $p = .015$. B: $t[565] = 2.33$, $p = .027$). Our female participants only agreed with the trend in the correspondence rating where the male virtual character with a high-pitched voice has higher audio-visual correspondence when paired with an ectomorphic body than when paired with a mesomorphic body ($t[565] = 2.31$, $p = .028$). Female participants only perceived a high-pitched voice with male ectomorphs to be marginally more believable than with mesomorphs ($t[565] = 2.02$, $p = .054$). However, we did not find statistically significant differences in the correspondence level when comparing a high-pitched voice paired with an ectomorphic body versus an endomorphic body.

Only male participants gave a statistically significant lower correspondence rating for male virtual characters with a medium-pitched voice when paired with the endomorphic body type than when paired with the mesomorphic body type ($t[565] = -2.35$, $p = .028$). Similarly, only male participants rated a low-pitched voice to be significantly more believable with endomorphs than ectomorphs ($t[565] = -2.07$, $p = .048$). However, we observed no statistically significant contrasts in the perceived believability ratings for medium-pitched voice on male virtual characters among the different body types.

Male virtual characters with a low-pitched voice got statistically significant lower perceived audio-visual correspondence and believability ratings when paired with an ectomorphic body than when paired with a mesomorphic body (AVC: male participants: $t[565] = -2.62$, $p = .012$; female participants: $t[565] = -2.07$, $p = .048$. B: male participants: $t[565] = -2.75$, $p = .008$; female participants: $t[565] = -2.30$, $p = .029$). Moreover, only male participants rated a low-pitched voice to have statistically significantly lower audio-visual correspondence and to be significantly more believable with endomorphs than with ectomorphs ($t[565] = -2.07$, $p = .048$).

4.3 Correlations

We computed the cumulative Pearson product-moment correlation coefficients between the perceived audio-visual correspondence and believability. We found strong correlations for the male ($r = .907$, $p < .001$) and female ($r = .909$, $p < .001$) virtual characters.

5 DISCUSSION

Overall, the results suggest that there were some statistically significant differences between the mean of the dependent variables, perceived audio-visual correspondence and believability, for each of the experimental treatment levels. We found that body type alone only affected the level of audio-visual correspondence and the perceived believability of male virtual characters. Meanwhile, voice pitch alone can affect the level of audio-visual correspondence [Kao et al. 2021] and the perceived believability of both female and male virtual characters. Moreover, the effect of voice pitch on our dependent variables (perceived audio-visual correspondence and believability) depended on the body type it was paired with and vice-versa. We observed strong correlations between the perceived audio-visual correspondence and the believability ratings of all the virtual characters. Finally, results from both male and female participants concluded with mostly the same main and interaction effects, with exceptions for extreme pairs that are discussed in the later part of this section.

Studies on humans showed that both female and male virtual characters with larger body and facial dimensions, classified as endomorphs, received higher audio-visual correspondence ratings and

were perceived as more believable when paired with low-pitched voices, and virtual characters with longer and smaller body and facial dimensions, classified as ectomorphs, to be associated with high-pitched voices [Cartei and Reby 2013; Macari et al. 2014]. In line with published human research, our participants rated the mesomorphic male virtual character with a medium-pitched voice to have the highest perceived audio-visual correspondence and to be the most believable among all the treatment levels for the male virtual character. Similarly, our results show significantly higher perceived audio-visual correspondence and believability ratings for endomorphic female virtual characters with a low-pitched voice than the other voice-pitch levels. In general, our participants perceived endomorphic virtual characters as the least believable with high-pitched voices than the other pitch levels. Other factors, such as the different facial width-to-height ratio (FWHR) in our virtual characters, may also influence our results. The endomorphic virtual characters we used in this study had the largest FWHR (widest faces) than the other body types. Published literature showed that wider faces in virtual characters and lower voice pitch were both perceived as more aggressive and dominant. This association may contribute to why our participants perceived endomorphs to be more believable with lower-pitched voices [Ferstl et al. 2021a; Jones et al. 2010]. Along the same trend, our participants perceived female mesomorphs and endomorphs with medium- and low-pitched voices as more believable with higher audio-visual correspondence, than when paired with the high-pitched voice. Our results for male virtual characters also show that participants agreed upon the general predicted trend where mesomorphs and endomorphs were less believable with less audio-visual correspondence when paired with high-pitched voices. However, there were no statistically significant differences when either the male mesomorph or the endomorph was paired with either a low- or a medium-pitched voice. Here, we speculated that our participants noticed the oddness of the mesomorphic body-high voice pitch and the endomorphic body-high voice pitch pairs. Correspondingly, a high-pitched voice was perceived as more believable and a better match with ectomorphs over the other two body types (**RQ1.1** & **RQ1.2**).

The overall quantitative data suggested that both male and female virtual characters were perceived as more believable, with higher audio-visual correspondence when paired with low- or medium-pitched voices when compared to the high-pitched voice (**RQ2.1** & **RQ2.2**). Aligning with our results, research on humans showed that a low-pitched voice for both male and female speakers was perceived as more dominant, and both men and women preferred low-pitched voices for men [Jones et al. 2010]. This suggested that virtual characters can be perceived as believable with a lower-than-average human voice pitch. Since our participants perceived the female ectomorphic virtual character to be the most believable with the medium-pitched voice, we think that the `F0` used for the high pitch level may be too high for virtual characters of this style.

Despite the commonest of the mesomorph body type [Koleva et al. 2002], our participants rated ectomorphs to have the overall highest level of perceived audio-visual correspondence and believability across the three body types (**RQ3.1** & **RQ3.2**). Moreover, our findings among both male and female participants agreed with that of Mitchell et al.'s [Mitchell et al. 2011], where they found the mismatch in voice and face to create discomfort and uncanniness. The strong correlation (**RQ4**) between our perceived audio-visual correspondence and believability ratings suggested that one does influence another.

Additionally (**RQ5**), our results show that male participants gave more extreme ratings for female virtual characters with combination pairs that opposed the published results from human research. Likewise, male participants were also more likely to give higher ratings for virtual characters with body type and voice pitch levels that were associated with human research. For example, only male participants gave lower perceived audio-visual correspondence and believability ratings for the ectomorphic female virtual character with a low-pitched voice. Moreover, our male participants perceived the ectomorphic female virtual character with a high-pitched voice to be more believable than our female participants did. Research showed that men had a stronger preference than women for women's voices with higher pitch [Jones et al. 2010]. Overall our male participants were more generous with their ratings. Furthermore, we observed that our participants noticed the extreme treatment pairs that were associated with human research [Macari et al. 2017; Pawelec et al. 2020], which included ectomorphs with high-pitched voices and endomorphs with low-pitched voices for the virtual characters of their sex. For example, only male participants rated male ectomorphs more believable when paired with a high-pitched voice than a low-pitched one. Similarly, although both participants sexes perceived the female endomorph with a low-pitched voice to be the most believable compared to the endomorph with other voice pitch levels, only female participants rated the female endomorph to have a marginally higher perceived audio-visual correspondence when paired with a low-pitched voice than when paired with a medium-pitched voice.

Although we expected the results to be the same for both female and male virtual characters because research showed that the pitch range for both sexes was similar [Traunmüller and Eriksson 1995], qualitative survey responses suggested that some of our participants believed that the acceptable voice pitch ranges were different for each sex due to the participants' different personal experiences or influences from current media. These participants did not specify how the ranges were different. A participant mentioned that some voice-body pairs appear to be off-putting, but not necessarily unbelievable because they felt like outliers did exist. Furthermore, we observed that the effect of participants' sex was only statistically significant in female virtual characters, which was a plausible bias caused by the majority of male participants in our pool.

From the optional survey responses, we discovered that our participants associated the treatment pairs with existing virtual characters from games, media, or real people they know. They often described similar personalities or roles. For example, several participants commented on the high-pitched voice for sounding too "juvenile" for the virtual character that resembled a "professional." Additionally, some participants mentioned that the virtual character looked and sounded "confident" and commented on the perceived emotions of the virtual character. The virtual characters were designed to have limited emotions to remain neutral but were perceived as tired or depressed. Even though the participants didn't directly mention whether those emotions affected the perceived audio-visual correspondence or believability scores, past studies

showed that emotions could affect voice pitch [Breitenstein et al. 2001].

Some participants mentioned that they wanted to edit the ratings they gave to the previous virtual character after seeing the other combinations. Whilst the treatment combinations were randomized, some participants commented that they felt like the voice belonged to the first virtual character they saw, resulting in a higher correspondence score for earlier pairs and difficulty in determining the "uniqueness" of the later voices. Literature showed that first impressions influenced whether humans build relations with others and find their interactions believable [Bergmann et al. 2012].

5.1 Limitations

Our results should be treated cautiously since this study was limited in various aspects. First, our study was conducted on participants who self-reported as English speakers who resided in the United States with American English-speaking virtual characters. Therefore, we cannot guarantee the same results for other non-US cultures. As discussed earlier, ratings were likely influenced by the participant's experience. Since people from different places vary by size, people of different cultures may have different thresholds for what they determine to be a larger or smaller virtual character. Furthermore, we cannot assume the same results for other languages, dialects, and accents, especially tonal languages, where pitch influences the meaning of the speech.

Multiple attributes characterize voices, but our study only focused on a single vocal feature: `F0`, perceived as pitch on a limited number of voices. Moreover, a virtual character is also characterized by various aspects that we kept constant in our study, including, but not limited to, age, style, personality, emotions, and clothing. Therefore, we do not know if the same results exist when other acoustic features and virtual character elements are also taken into account.

Finally, this study was highly dependent on the voice synthesis method used to manipulate the voice pitch and the animation techniques used to create lip-sync animation. A few participants commented that sometimes the virtual characters felt robotic, and the lip-syncs were slightly off due to the video buffering. Thus, these technical limitations may have affected our results.

6 CONCLUSIONS AND FUTURE WORK

In conclusion, our results suggest that the virtual character's perceived audio-visual correspondence level, that is, the match between the virtual character's physical dimensions from head to body and the virtual character's voice pitch, influenced the virtual character's perceived believability. It was shown that human audiences had expectations for the appropriate voice pitch that belonged to certain body types. Thus, when virtual characters of a certain body type had a voice pitch that matched the viewer's expectation for how that body should sound, the virtual character became more believable. These expectations were drawn from real-life experiences that aligned with the general consensus shown in human studies. However, the perceived audio-visual correspondence in virtual characters may not be sufficient in determining the perceived believability of the virtual characters since there are other contributors, such as emotion, personality, and interactions.

For future research, we would like to strengthen the impact of our project by exploring other factors that were the limitations of this study. We would like to explore the voice-to-body relationship in different cultures and languages because virtual characters are often exposed to audiences worldwide. Additionally, since culture also influences people's preference on body type [Sewell 2011], another possible future work is to assess whether different levels of attractiveness can influence the levels of believability. Furthermore, we would like to explore the use of virtual characters to increase diversity and inclusion in online classrooms. We would also like to investigate the same relationship across different virtual character styles, emotions, and settings. Although we specified that the perceived believability in our study is the likeliness that the virtual characters produce the give voice, our ratings still depended on what the participants perceived as "believable" to some extent. Therefore, we would like to extend to the other contexts of perceived believability, such as interactivity [Bogdanovych et al. 2016], and present our treatment combinations in a more immersive, engaging, and interactive experience. One possibility is to directly test the effects of perceived believability, such as the virtual character's ability to influence the participants [Bogdanovych et al. 2016].

ACKNOWLEDGMENTS

This project was supported by an NSF RETTL Award (award number: 2201019).

REFERENCES

[n. d.]. Hair Color by Country 2022. https://worldpopulationreview.com/country-rankings/hair-color-by-country

2009. *Handbook of Multimedia for Digital Entertainment and Arts*. Springer US. https://doi.org/10.1007/978-0-387-89024-1

2017. The UW/NU Corpus. http://depts.washington.edu/phonlab/resources/uwnu2/

2019. Configuring the Pitch Contour. https://www.fon.hum.uva.nl/praat/manual/Intro_4_2__Configuring_the_pitch_contour.html

J. Abitbol, P. Abitbol, and B. Abitbol. 1999. Sex hormones and the female voice. *Journal of Voice* 13, 3 (Sept. 1999), 424–446. https://doi.org/10.1016/s0892-1997(99)80048-4

E. P Altenberg and C. T Ferrand. 2006. Fundamental frequency in monolingual English, bilingual English/Russian, and bilingual English/Cantonese young adult women. *Journal of Voice* 20, 1 (2006), 89–96.

C.D. Aronovitch. 1976. The Voice of Personality: Stereotyped Judgments and their Relation to Voice Quality and Sex of Speaker. *The Journal of Social Psychology* 99, 2 (1976), 207–220. https://doi.org/10.1080/00224545.1976.9924774 arXiv:https://doi.org/10.1080/00224545.1976.9924774 PMID: 979189.

M.P. Aylett, A. Vinciarelli, and M. Wester. 2017. Speech synthesis for the generation of artificial personality. *IEEE transactions on affective computing* 11, 2 (2017), 361–372.

S. P Badathala, N. Adamo, N. J Villani, and H. N Dib. 2018. The effect of gait parameters on the perception of animated agents' personality. In *International Conference on Augmented Reality, Virtual Reality and Computer Graphics*. Springer, 464–479.

R. J Baken and R. F Orlikoff. 2000. *Clinical measurement of speech and voice*. Cengage Learning.

A. M. Baranowski and H. Hecht. 2018. Effect of camera angle on perception of trust and attractiveness. *Empirical Studies of the Arts* 36, 1 (2018), 90–100.

B. Barsties, R. Verfaillie, P. Dicks, and Y. Maryn. 2016. Is the speaking fundamental frequency in females related to body height? *Logopedics Phoniatrics Vocology* 41 (2016), 27–32. Issue 1. https://doi.org/10.3109/14015439.2014.941928

P. Belin, S. Fecteau, and C. Bedard. 2004. Thinking the voice: neural correlates of voice perception. *Trends in cognitive sciences* 8, 3 (2004), 129–135.

Kirsten Bergmann, Friederike Eyssel, and Stefan Kopp. 2012. A second chance to make a first impression? How appearance and nonverbal behavior affect perceived warmth and competence of virtual agents over time. In *Intelligent Virtual Agents: 12th International Conference, IVA 2012, Santa Cruz, CA, USA, September, 12-14, 2012. Proceedings 12*. Springer, 126–138.

M. Berry, S. Lewin, and S. Brown. 2022. Correlated expression of the body, face, and voice during character portrayal in actors. *Scientific Reports* 12 (2022). Issue 1. https://doi.org/10.1038/s41598-022-12184-7

Anton Bogdanovych, Tomas Trescak, and Simeon Simoff. 2016. What makes virtual agents believable? *Connection Science* 28, 1 (2016), 83–108.

S. Bommarito. 2019. Correlation Between Voice, Speech, Body and Facial Types in Young Adults. *Global Journal of Otolaryngology* 20 (2019). Issue 4. https://doi.org/10.19080/gjo.2019.20.556041

C. Breitenstein, D. V Lancker, and I. Daum. 2001. The contribution of speech rate and pitch variation to the perception of vocal emotions in a German and an American sample. *Cognition & Emotion* 15, 1 (2001), 57–79.

J.P. Cabral, B.R. Cowan, K. Zibrek, and R. McDonnell. 2017. The influence of synthetic voice on the evaluation of a virtual character. *Proceedings of the Annual Conference of the International Speech Communication Association* 2017-August, 229–233. https://doi.org/10.21437/Interspeech.2017-325

V. Cartei and D. Reby. 2013. Effect of formant frequency spacing on perceived gender in pre-pubertal children's voices. *PLoS ONE* 8 (2013). Issue 12. https://doi.org/10.1371/journal.pone.0081022

JE L. Carter and B. H. Heath. 1990. *Somatotyping: development and applications*. Vol. 5. Cambridge university press.

F. Christie and V. Bruce. 1998. The role of dynamic information in the recognition of unfamiliar faces. *Memory & cognition* 26, 4 (1998), 780–790.

R. O. Coleman. 1976. A Comparison of the Contributions of Two Voice Quality Characteristics to the Perception of Maleness and Femaleness in the Voice. *Journal of Speech and Hearing Research* 19, 1 (1976), 168–180. https://doi.org/10.1044/jshr.1901.168 arXiv:https://pubs.asha.org/doi/pdf/10.1044/jshr.1901.168

S.A. Collins and C. Missing. 2003. Vocal and visual attractiveness are related in women. *Animal behaviour* 65, 5 (2003), 997–1004.

L. Bernadete Rocha de Souza and M. Marques dos Santos. 2018. Body mass index and acoustic voice parameters: is there a relationship? *Brazilian Journal of Otorhinolaryngology* 84 (2018), 410–415. Issue 4. https://doi.org/10.1016/j.bjorl.2017.04.003

E. D'haeseleer, H. Depypere, S. Claeys, N. Baudonck, and K. Van Lierde. 2012. The impact of hormone therapy on vocal quality in postmenopausal women. *Journal of Voice* 26, 5 (2012), 671–e1.

Patrick Doyle. 2002. Believability through context using" knowledge in the world" to create intelligent characters. In *Proceedings of the first international joint conference on Autonomous agents and multiagent systems: part 1*. 342–349.

R EBU-Recommendation. 2011. Loudness normalisation and permitted maximum level of audio signals.

A. Edwards and C. Newell. 2005. Lively voice: a new model for speaking synthetic characters How changing physical stiffness parameters of virtual objects alter our perception of roughness during force feedback based haptic exploration View project.

J.T. Eichhorn, R. D. Kent, D. Austin, and H. K. Vorperian. 2018. Effects of Aging on Vocal Fundamental Frequency and Vowel Formants in Men and Women. *Journal of Voice* 32 (2018), 644.e1–644.e9. Issue 5. https://doi.org/10.1016/j.jvoice.2017.08.003

S. Evans, N. Neave, and D. Wakelin. 2006. Relationships between vocal characteristics and body size and shape in human males: An evolutionary explanation for a deep male voice. *Biological Psychology* 72 (5 2006), 160–163. Issue 2. https://doi.org/10.1016/j.biopsycho.2005.09.003

Ylva Ferstl, Michael McKay, and Rachel McDonnell. 2021a. Facial feature manipulation for trait portrayal in realistic and cartoon-rendered characters. *ACM Transactions on Applied Perception (TAP)* 18, 4 (2021), 1–8.

Y. Ferstl, S. Thomas, C. Guiard, C. Ennis, and R. McDonnell. 2021b. Human or Robot? Investigating voice, appearance and gesture motion realism of conversational social agents. In *Proceedings of the 21st ACM international conference on intelligent virtual agents*. 76–83.

W T. Fitch and J. Giedd. 1999. Morphology and development of the human vocal tract: A study using magnetic resonance imaging. *The Journal of the Acoustical Society of America* 106, 3 (1999), 1511–1522.

K. L Garrett and E C. Healey. 1987. An acoustic analysis of fluctuations in the voices of normal adult speakers across three times of day. *The Journal of the Acoustical Society of America* 82, 1 (1987), 58–62.

Marylou Pausewang Gelfer and Victoria A. Mikos. 2005. The Relative Contributions of Speaking Fundamental Frequency and Formant Frequencies to Gender Identification Based on Isolated Vowels. *Journal of Voice* 19, 4 (2005), 544–554. https://doi.org/10.1016/j.jvoice.2004.10.006

J. Grzybowska and S. Kacprzak. 2016. Speaker Age Classification and Regression Using i-Vectors.. In *INTERSPEECH*. 1402–1406.

I. Guimarães and E. Abberton. 2005. Health and voice quality in smokers: an exploratory investigation. *Logopedics Phoniatrics Vocology* 30, 3-4 (2005), 185–191.

H. Hatano, T. Kitamura, H. Takemoto, P. Mokhtari, K. Honda, and S. Masaki. 2012. Correlation between vocal tract length, body height, formant frequencies, and pitch frequency for the five Japanese vowels uttered by fifteen male speakers. *13th Annual Conference of the International Speech Communication Association 2012* 1, 402–405. https://doi.org/10.21437/interspeech.2012-143

D. Higgins, K. Zibrek, J. Cabral, D. Egan, and R. McDonnell. 2022. Sympathy for the digital: Influence of synthetic voice on affinity, social presence and empathy for photorealistic virtual humans. *Computers and Graphics (Pergamon)* 104 (2022), 116–128. https://doi.org/10.1016/j.cag.2022.03.009

Susan Hughes, Marissa A Harrison, and Gordon G Gallup. [n. d.]. SEX-SPECIFIC BODY CONFIGURATIONS CAN BE ESTIMATED FROM VOICE SAMPLES. , 343–355 pages. Issue 4.

F. Joassin, M. Pesenti, P. Maurage, E. Verreckt, R. Bruyer, and S. Campanella. 2011. Cross-modal interactions between human faces and voices involved in person recognition. *Cortex* 47 (3 2011), 367–376. Issue 3. https://doi.org/10.1016/j.cortex.2010.03.003

Benedict C Jones, David R Feinberg, Lisa M DeBruine, Anthony C Little, and Jovana Vukovic. 2010. A domain-specific opposite-sex bias in human preferences for manipulated voice pitch. *Animal Behaviour* 79, 1 (2010), 57–62.

Jessica Junger, Katharina Pauly, Sabine Bröhr, Peter Birkholz, Christiane Neuschaefer-Rube, Christian Kohler, Frank Schneider, Birgit Derntl, and Ute Habel. 2013. Sex matters: Neural correlates of voice gender perception. *NeuroImage* 79 (2013), 275–287. https://doi.org/10.1016/j.neuroimage.2013.04.105

M. Kamachi, H. Hill, K. Lander, and E. Vatikiotis-Bateson. 2003. Putting the face to the voice': Matching identity across modality. *Current Biology* 13, 19 (2003), 1709–1714.

Dominic Kao, Rabindra Ratan, Christos Mousas, Amogh Joshi, and Edward F. Melcer. 2022. Audio Matters Too: How Audial Avatar Customization Enhances Visual Avatar Customization. *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3491102.3501848

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. Issue CHIPLAY. https://doi.org/10.1145/3474665

M Koleva, A Nacheva, and M Boev. 2002. Somatotype and disease prevalence in adults. *Reviews on environmental health* 17, 1 (2002), 65–84.

J. Kreiman and D. Sidtis. 2011. *Foundations of voice studies*. Wiley-Blackwell, Chichester, England.

A. Kuznetsova, P. B Brockhoff, and R. HB Christensen. 2017. lmerTest package: tests in linear mixed effects models. *Journal of statistical software* 82 (2017), 1–26.

L. Lachs and D.B Pisoni. 2004. Crossmodal source identification in speech perception. *Ecological Psychology* 16, 3 (2004), 159–187.

K. Lightstone, R. Francis, and L. Kocum. 2011. University faculty style of dress and students' perception of instructor credibility. *International Journal of Business and Social Science* 2, 15 (2011).

A Bryan Loyall, Joseph Bates, Jill Fain Lehman, Tom Mitchell, and Nils Nilsson. 1997. Believable Agents: Building Interactive Personalities.

H.H. Lu, S.E. Weng, Y.F. Yen, H.H Shuai, and W.H. Cheng. 2021. Face-based Voice Conversion: Learning the Voice behind a Face. *MM 2021 - Proceedings of the 29th ACM International Conference on Multimedia*, 496–505. https://doi.org/10.1145/3474085.3475198

A. T. Macari, I. A. Karam, D. Tabri, D. Sarieddine, and A.L. Hamdan. 2014. Correlation between the length and sagittal projection of the upper and lower jaw and the fundamental frequency. *Journal of Voice* 28 (2014), 291–296. Issue 3. https://doi.org/10.1016/j.jvoice.2013.10.003

A. T. Macari, I. A. Karam, D. Tabri, D. Sarieddine, and A.L. Hamdan. 2015. Formants frequency and dispersion in relation to the length and projection of the upper and lower jaws. *Journal of Voice* 29 (2015), 83–90. Issue 1. https://doi.org/10.1016/j.jvoice.2014.05.011

A. T. Macari, I. A. Karam, G. Ziade, D. Tabri, D. Sarieddine, E.S. Alam, and A.L. Hamdan. 2017. Association Between Facial Length and Width and Fundamental Frequency. *Journal of Voice* 31 (2017), 410–415. Issue 4. https://doi.org/10.1016/j.jvoice.2016.12.001

C. Maguinness, C. Roswandowitz, and K. von Kriegstein. 2018. Understanding the mechanisms of familiar voice-identity recognition in the human brain. *Neuropsychologia* 116 (2018), 179–193.

S. Marsella and J. Gratch. 2003. Modeling coping behavior in virtual humans: don't worry, be happy. In *Proceedings of the second international joint conference on Autonomous agents and multiagent systems*. 313–320.

L. W. Mavica and E. Barenholtz. 2013. Matching voice and face identity from static images. *Journal of Experimental Psychology: Human Perception and Performance* 39 (2013), 307–312. Issue 2. https://doi.org/10.1037/a0030945

W.J. Mitchell, K.A. Szerszen, A.S. Lu, P.W. Schermerhorn, M. Scheutz, and K.F. MacDorman. 2011. A mismatch in the human realism of face and voice produces an uncanny valley. , 10–12 pages. Issue 1. https://doi.org/10.1068/i0415

R. Mondragón-Ceballos, M. D. G. Granados, A. L. Cerda-Molina, R. Chavira-Ramírez, and L. E. Hernández-López. 2015. Waist-to-hip ratio, but not body mass index, is associated with testosterone and estradiol concentrations in young women. *International Journal of Endocrinology* 2015 (2015). https://doi.org/10.1155/2015/654046

R. Niewiadomski and C. Pelachaud. 2011. How Is Believability of a Virtual Agent Related to Warmth, Competence, Personification, and Embodiment? , 431–448 pages. Issue 5.

T. Oh, T. Dekel, C. Kim, I. Mosseri, W.T. Freeman, M. Rubinstein, and W. Matusik. 2019. Speech2Face: Learning the Face Behind a Voice. (2019).

L.P. Pawelec, K. Graja, and A. Lipowicz. 2020. Vocal Indicators of Size, Shape and Body Composition in Polish Men. *Journal of Voice* (2020). https://doi.org/10.1016/j.jvoice.2020.09.011

Katarzyna Pisanski, Paul J. Fraccaro, Cara C. Tigue, Jillian J.M. O'Connor, Susanne Röder, Paul W. Andrews, Bernhard Fink, Lisa M. DeBruine, Benedict C. Jones, and David R. Feinberg. 2014. Vocal indicators of body size in men and women: a meta-analysis. *Animal Behaviour* 95 (Sept. 2014), 89–99. https://doi.org/10.1016/j.anbehav.2014.06.011

K. Pisanski, B. C. Jones, B. Fink, J. J.M. O'Connor, L. M. DeBruine, S. Röder, and D. R. Feinberg. 2016. Voice parameters predict sex-specific body morphology in men and women. *Animal Behaviour* 112 (2016), 13–22. https://doi.org/10.1016/j.anbehav.2015.11.008

T. Rakić, M.C. Steffens, and A. Mummendey. 2011. Blinded by the Accent! The Minor Role of Looks in Ethnic Categorization. *Journal of personality and social psychology* 100, 1 (2011), 16–29.

Mark O Riedl and Andrew Stern. 2006. Failing believably: Toward drama management with autonomous actors in interactive narratives. In *Technologies for Interactive Digital Storytelling and Entertainment: Third International Conference, TIDSE 2006, Darmstadt, Germany, December 4-6, 2006. Proceedings 3*. Springer, 195–206.

Rachel Sewell. 2011. What is appealing? sex and racial differences in perceptions of the physical attractiveness of women. (2011).

D. R. R. Smith, R. D. Patterson, R. Turner, H. Kawahara, and T. Irino. 2005. The processing and perception of size information in speech sounds. *The Journal of the Acoustical Society of America* 117 (2005), 305–318. Issue 1. https://doi.org/10.1121/1.1828637

H. M.J. Smith, A. K. Dunn, T. Baguley, and P. C. Stacey. 2016. Matching novel face and voice identity using static and dynamic facial images. *Attention, Perception, and Psychophysics* 78 (2016), 868–879. Issue 3. https://doi.org/10.3758/s13414-015-1045-8

F. Thomas and O. Johnston. 1981. The Illusion of life: Disney animation in New York. *NY Hyperion* (1981).

S. Thomas, Y. Ferstl, R. Mcdonnell, and C. Ennis. [n. d.]. Investigating How Speech And Animation Realism Influence The Perceived Personality Of Virtual Characters And Agents.

A. Tinwell, M. Grimshaw-Aagaard, and A. Williams. 2010. Uncanny behaviour in survival horror games. *Games Computing and Creative Technologies: Journal Articles (Peer-Reviewed)* 2 (05 2010). https://doi.org/10.1386/jgvw.2.1.3_1

Hartmut Traunmüller and Anders Eriksson. 1995. The frequency range of the voice fundamental in the speech of male and female adults. *Unpublished manuscript* 11 (1995).

D.G. Walshe, E.J. Lewis, S.I. Kim, K. O'Sullivan, and B.K. Wiederhold. 2003. Exploring the use of computer games and virtual reality in exposure therapy for fear of driving following a motor vehicle accident. *CyberPsychology & Behavior* 6, 3 (2003), 329–334.

Taiba Majid Wani, Teddy Surya Gunawan, Syed Asif Ahmad Qadri, M. Kartiwi, and E. Ambikairajah. 2021. A Comprehensive Review of Speech Emotion Recognition Systems. , 47795–47814 pages. https://doi.org/10.1109/ACCESS.2021.3068045

P. Wisessing, K. Zibrek, D. W. Cunningham, J. Dingliana, and R. McDonnell. 2020. Enlighten Me: Importance of Brightness and Shadow for Character Emotion and Appeal. *ACM Trans. Graph.* 39, 3, Article 19 (4 2020), 12 pages. https://doi.org/10.1145/3383195

S. Wuhrer and C. Shu. 2013. Estimating 3D human shapes from measurements. *Machine Vision and Applications* 24 (8 2013), 1133–1147. Issue 6. https://doi.org/10.1007/s00138-012-0472-y

A. Yamauchi, H. Imagawa, H. Yokonishi, K.I. Sakakibara, and N. Tayama. 2022. Gender- and Age- Stratified Normative Voice Data in Japanese-Speaking Subjects: Analysis of Sustained Habitual Phonations. *Journal of Voice* (2022). https://doi.org/10.1016/j.jvoice.2021.12.002

A.W. Young, S. Frühholz, and S.R. Schweinberger. 2020. Face and voice perception: Understanding commonalities and differences. *Trends in Cognitive Sciences* 24, 5 (2020), 398–410.

Z. Zhang. 2016. Mechanics of human voice production and control. *The Journal of the Acoustical Society of America* 140 (2016), 2614–2635. Issue 4. https://doi.org/10.1121/1.4964509

Z. Zhang, B. Wu, and B. Schuller. 2019. Attention-augmented end-to-end multi-task learning for emotion prediction from speech. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 6705–6709.

R. I Zraick, M. A Gentry, L. Smith-Olinde, and B. A Gregg. 2006. The effect of speaking context on elicitation of habitual pitch. *Journal of Voice* 20, 4 (2006), 545–554.

A APPENDIX: QUESTIONNAIRES

Table 1: Questions measuring the perceived believability of the virtual character.

#	Question	View
Q1	Rate the believability of the virtual character's voice and appearance combined.	full-body
Q2	Rate how well you agree with the following: I would expect the virtual character to sound like this from the way they look.	full-body
Q3	Rate how well you agree with the following: I would expect the virtual character to look like this from the way they sound.	full-body

Table 2: Questions measuring the perceived audio-visual correspondence in the virtual character.

#	Question	View
Q1	How well does the virtual character's voice match their face?	close-up
Q2	How well does the virtual character's voice match their body?	full-body
Q3	How appropriate were both the visual and audial characteristics combined of the virtual character?	full-body
Q4	Rate how well you agree with the following: The virtual character's voice matches the virtual character's appearance.	full-body
Q5	Rate how well you agree with the following: The voice sounds unique to the virtual character.	full-body

Table 3: Free-form questions.

#	Question	View
Q1	Any comments on the virtual characters' voices?	full-body
Q2	Any comments on the virtual characters' appearances?	full-body
Q3	Any comments on the virtual characters' appearance-voice match?	full-body