Uncanny Valley Effect is Amplified with Multimodal Stimuli and Varies Across Ages

Emre Yorgancıgil a*, Burcu A. Urgen b,c,d, Funda Yıldırım e,f

a Neuroscience PhD Program, Institute of Health Sciences, Yeditepe University, 26 August Campus, Kayışdağı 34755, İstanbul – Turkey

b Department of Psychology, Bilkent University, 06800, Bilkent, Ankara-Turkey

c Neuroscience Graduate Program, Bilkent University, 06800, Bilkent, Ankara-Turkey

d Aysel Sabuncu Brain Research Center and National Magnetic Resonance Research Center (UMRAM), Ankara-Turkey

e Computational Neuroscience and Visual Perception Laboratory, Department of Computer Engineering, Yeditepe University, 26 August Campus, Kayışdağı 34755, İstanbul – Turkey

f Cognitive Science Master’s Program, Yeditepe University, 26 August Campus, Kayışdağı 34755, İstanbul – Turkey

Abstract

“Uncanny Valley Hypothesis” suggests that humanoid objects that materialize human beings virtually but not entirely realistically may elicit uncanny feelings of eeriness and revulsion in observers. While the uncanny valley (UV) has been largely investigated with a focus on the visual aspects of the robot-like designs with young adults, the auditory components that may contribute to this effect and how visual and auditory factors jointly play a role in uncanny reports across different generations has not been examined. In the present study, we investigated how multimodal stimuli and the congruence of visual and auditory aspects of the stimuli contribute to the uncanniness perception and differ from the audio and visual components across generations. Young and old adults rated animations that were presented in audio-visual, audio-only and visual-only modalities in terms of uncanniness. The visual and auditory aspects of the stimuli had four levels of naturalness: robot (unrealistic), semi-robot (semi-realistic), human-like (realistic) and human (real). Our results show that audio-visual stimuli have an amplified effect on UV scores than only auditory and only visual stimuli. In addition, multimodal stimuli that have incongruent audio and visual components elicited significantly higher uncanny scores than stimuli that have congruent components. However, the difference between congruent and incongruent stimuli
were more pronounced in the younger group compared to the older group. We also found that younger generations are more sensitive to naturalness layers of audio-visual stimuli than older generations. In conclusion, uncanny valley effect is modulated by stimulus modality, congruence of visual and auditory modalities, naturalness as well as age.

**Keywords:** Uncanny valley, human-robot interaction, audio-visual perception, face-voice mismatch

### 1. Introduction

Masahito Mori’s “Uncanny Valley Hypothesis” suggests that humanoid objects that materialize human beings virtually but not entirely realistically may elicit uncanny feelings of eeriness and revulsion in observers. He used a hypothetical curve to describe this relationship and indicated a sharp dip in this curve at almost humanlike levels as an ‘uncanny’ valley (Mori et al., 2012). In the last two decades this issue has gained more significance (Perry, 2014).

Visual components carry important clues for 3D graphical design and animations. A recent study suggested that the gender, facial asymmetry, skin colour, hair colour and realism level preserved by a virtual face are important factors determining the uncanny valley effect (Schwind et al., 2015). Also, smooth and natural skin colours were evaluated as more likeable and attractive by the subjects. Realistic face designs were associated with higher affinity and trustworthiness compared to cartoon-like face designs (seymour et al., 2019). Expectation might also play an important role in evaluating humanoid faces. Atypical facial features like different shape and place of eyes and inconsistent realism levels may cause perceptual mismatch in subjects. This mismatch situation can reveal uncanny valley feelings (Kätsyri et al., 2015; Seyama & Nagayama, 2007).

Among the sensory effects, auditory effects are lesser known compared to visual effects and the choices regarding robot voices are harder to make for improving human-robot interaction. Components like pitch, pitch range, volume, and speech rate have a role on human perception of voice (Dou et al., 2019). Like a validation of our natural tendencies, people have more positive attitudes toward human-like voices over synthetic voices. There has been a correlation between positive emotions and naturalness level of voice (Stevens et al., 2005). Also, in phone communication, people prefer human voice over robot voice (Drager et al., 2004). During the communication with a robot, people consider the robot’s face and voice together. When talking with a mechanical looking robot, they expect a synthesized voice from the robot. When there is a mismatch between voice and face of an agent, cognitive dissonance may occur due to violation of expectations about human/robot categorization (D. G. V. Mitchell, 2011). It might be possible that the audio-visual uncanny valley effect may take place due to this mismatch effect (Meah & Moore,
In gaming, asynchronous lip movements and voice of characters contribute to the uncanny valley effect (Tinwell et al., 2010). Also, inconsistency between voice and social identity, gender and accent can create undesirable emotions (Gong & Nass, 2007).

If the uncanny valley effect is due to the termination of expectancy, people should find matching audio and visuals less uncanny, even if they are perceived more disturbing. On the other hand, if the uncanny valley effect is an outcome depending solely on the level of unnaturalness, then regardless of the congruency between visual and auditory aspects, disturbing unimodal sensory stimuli should determine the level of uncanniness. In the light of these associations found between the visual and auditory characteristics of digital character designs, we aim to explore how different levels of human-likeness in audio-visual characteristics of humanoid and robot faces will create an uncanny valley effect. We will explore unimodal conditions for the face and for the voice of the characters as well as multimodal face - voice stimuli to understand their effect on uncanny feelings.

Age can be a potent factor for the uncanny valley effect. A body of human-robot interaction studies compare how different generations react to robot designs and robotic behaviours (Langer & Levy-Tzedek, 2020). Children can feel discomfort with humanlike robots and uncanny valley effect could emerge with improved mental capacity in older children (Brink et al., 2019; Strait et al., 2019). Although there has been an increased interest in understanding uncanny valley from a developmental perspective including children, there are much less studies involving older generations (Tung, 2016). A handful of studies show that elderly people have more negative attitudes to assistive technology and devices. For instance, older adults feel warmer emotions to humans than robots and prefer interacting with robots or humanoids only in consistent appearance (Destephe et al., 2015; Prakash & Rogers, 2013).

The aim of the present study is to test the individual and joint contribution of visual and auditory modalities on the uncanny reports in two different age groups. To this end, we investigated how the facial features (visual), the voice (auditory) and a combination of the two influence the uncanny reports. We used computer animations that vary in terms of naturalness in both visual and auditory modalities.

2. Materials and Methods

2.1. Participants

60 healthy participants participated in the study. Participants were divided into two age groups which will be referred to as the “older age group” and the “younger age group”. Each group had 30 participants. The older age group consisted of 13 women and 17 men which had an age range of 40-64 (M = 51.2, SD = 7.39), while the younger age group consisted of 12 women and 18 men which had an age range of 18-27 (M = 20.3, SD = 2.07). All members of the older age group were university graduates while all members of the younger age group were undergraduate and graduate students who reported to use computers in their daily lives. The study was conducted in compliance with the WMA Declaration of Helsinki, and participants
signed a consent form before the experiment. Participants did not receive any form of payment.

2.1. Experimental Stimuli
We used 16 different stimuli in our experiment. These stimuli were presented as supplementary material. These include 6 unimodal stimuli (3 only visual stimuli, 3 only auditory stimuli) and 10 multimodal stimuli. Only visual and only auditory stimuli were combined with each other to create 9 multimodal audio-visual stimuli. There was also one control multimodal stimulus which adds up to a total of 10 multimodal audio-visual stimuli (Table 1).

<table>
<thead>
<tr>
<th>Unimodal</th>
<th>Multimodal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Only Visual</td>
<td>Audio-visual</td>
</tr>
<tr>
<td>Only Auditory</td>
<td>Real Face (control stimulus)</td>
</tr>
<tr>
<td>Only Realistic</td>
<td>Realistic Face - Realistic Voice</td>
</tr>
<tr>
<td>Face</td>
<td>Realistic Face - Semi-realistic Voice</td>
</tr>
<tr>
<td>Only Unrealistic</td>
<td>Semi-realistic Face - Semi-realistic Voice</td>
</tr>
<tr>
<td>Face</td>
<td>Semi-realistic Face - Realistic Voice</td>
</tr>
<tr>
<td>Only Semi-realistic Face</td>
<td>Semi-realistic Face - Unrealistic Voice</td>
</tr>
<tr>
<td>Only Unrealistic</td>
<td>Unrealistic Face - Unrealistic Voice</td>
</tr>
<tr>
<td>Face</td>
<td>Unrealistic Face - Semi-realistic Voice</td>
</tr>
</tbody>
</table>

2.1.1. Visual Stimuli
As visual stimuli, we created animated faces at different levels of naturalness in the human-robot spectrum. These animations were modelled from a real human face (Figure 1a) and had three degrees: a lifelike humanoid design that we call the “realistic human face”, a primitive robot design that we call the “unrealistic human face”, and a face in between that we call the “semi-realistic human face” (Figure 1b-d).

The 3D models of faces were created using the Reallusion Character Creator 3® software. Default 3D humanlike templates were altered to create the models used in the images and video clips. The degree of the naturalness of the faces was manipulated by altering the texture of the underlying mesh, giving surface material a glossier look. The facial animation was created using the Reallusion iClone-7® software. Mouth movements were adjusted to follow the speech simultaneously by using a software’s text-to-speech plugin. Any lack of harmony between the speech and the video were fixed manually.

The real human face stimulus was created by recording a 33-year-old male model. He had no visible body hair except his eyebrows (Figure 1a). In the realistic human face all facial features imitated a real human face with proper portions and locations. It had no body hair except eyebrows and the skin colour was naturalistic.
The semi-realistic human face had light blue eyes, sunken cheeks, apparent but colourless lips, visible ears and nose. In contrast to the realistic human face, its skin colour was white. Due to this contrast between skin colour and facial features, we expected a mild uncanniness effect from this design (Figure 1c). The unrealistic human face had very basic facial features. The eye area had no details, the skin colour was dark grey, and the mouth, the nose and the ears were simply outlined. The surface material was metallic, bright and reflective to make the look compatible with basic robot designs (Figure 1d).

2.1.2. Auditory Stimuli
We recorded the voice of a 19-year-old, native English male speaker while he said the phrase “Hello. How are you doing?” in a sound recording studio. This sound file was used as the realistic voice condition in our study. In addition, this sound file was modified to manipulate the naturalness of the sound in different degrees: a robotic sound that we call the unrealistic voice and an artificial sound in between the human and robotic sounds that we call the semi-realistic voice. Pitch and formant parameters of the sound were adjusted (realistic voice: 3rd minor -80 formant -1.25, semi-realistic voice: 3rd minor -85 -1.25 formant, unrealistic voice: 2 minor 5 9.48 formant) to obtain the deserved effects. All sounds had a frequency interval between 60 and 512 Hz.

2.1.3. Audio-visual Stimuli
9 audio-visual video stimuli were created by pairing all three animated visual stimuli with all three types of auditory stimuli using Adobe Premiere®. There was an additional audio-visual stimulus, namely the ‘real’ audio-visual stimulus which consisted of the real human face paired with the realistic voice. This was used as a control condition.
2.2. Procedure

Participants were seated in a chair in front of a 17-inch LCD screen (Samsung SyncMaster 923NW LCD monitor, 1440x900 pixel, 60 Hz frequency) in a dark room. The screen height was set to the eye level of the participants, and the distance of the participant from the screen was 45 cm (Föcker et al., 2018; Hwang & Lee, 2021).

The experiment consisted of three types of blocks: 6 blocks of the visual stimuli, 6 blocks of the auditory stimuli, and 20 blocks of the audio-visual stimuli, presented in this order. In each visual stimuli block, the realistic, semi-realistic, and unrealistic human faces were shown twice in random order. In each auditory stimuli block, the realistic, semi-realistic, and unrealistic human voice stimuli were presented twice in random order. In each audio-visual block, the 10 audio-visual stimuli described in Table 1 were presented twice in random order. The 20 audio-visual blocks were presented as two sets of 10 blocks, separated by a 18-sec rest.

Each trial of the experiment within each block started with a fixation screen (1.5 sec), followed by the stimulus screen (4 sec). Upon viewing the stimuli, participants were asked to rate the uncanniness of the stimuli within 6 seconds (Figure 2). Participants rated the stimuli using a 1-to-9 point Likert scale where 1 indicated “normal” and 9 indicated “uncanny” (Turkish translations of the words were also placed in the rating screen). The experiment was programmed using the open source PsychoPy3 software, a Python library which has a special builder interface to conduct behavioural experiments (Peirce, 2009).

![Experimental design](image_url)

**Figure 2.** Experimental design: (a) Block structure of the whole experiment, b) The structure of a single trial.
2.3. Statistical Analysis

We computed the mean rating score, i.e. uncanniness for each type of stimuli in the experiment. Multivariate ANOVA analyses were employed to unimodal and multimodal stimuli separately to evaluate uncanniness ratings of stimuli between age groups.

Then we performed several ANOVAs to test our hypotheses. These include:

- A 2 (Age: Old, Young) x 3 (Naturalness: Realistic, Semi-realistic, Unrealistic) mixed ANOVA on the ratings given for the visual stimuli.
- A 2 (Age: Old, Young) x 3 (Naturalness: Realistic, Semi-realistic, Unrealistic) mixed ANOVA on the ratings given for the auditory stimuli.
- A 2 (Age: Old, Young) x 3 (Visual Stimuli: Realistic, Semi-realistic, Unrealistic) x 3 (Audio Stimuli: Realistic, Semi-realistic, Unrealistic) mixed ANOVA on the ratings of the audio-visual stimuli. This ANOVA was performed to check whether the congruence of the audio and visual stimuli has an effect of uncanny ratings.
- A 2 (Age: Old, Young) x 2 (Congruency: Congruence, Incongruence) mixed ANOVA on the ratings of the congruency situation. This ANOVA was performed to check at what level congruent and incongruent face - voice pairs differ from each other.
- An omnibus 2 (Age: Old, Young) x 3 (Modality: Only Visual, Only Auditory, Audio-visual) x 3 (Naturalness: Realistic, Semi-realistic, Unrealistic) mixed ANOVA. In this ANOVA, the realistic audio-visual condition was defined as the mean rating of all conditions that had a realistic human face (i.e. realistic face - realistic voice, realistic face - semi-realistic voice, and realistic face - unrealistic voice); the semi-realistic audio-visual condition was defined as the mean rating of all conditions that had a semi-realistic human face (i.e. semi-realistic face - realistic voice, semi-realistic face - semi-realistic voice, and semi-realistic face - unrealistic voice); and the unrealistic audio-visual condition was defined as the mean rating of all conditions that had an unrealistic human face (i.e. unrealistic face - realistic voice, unrealistic face - semi-realistic voice, and unrealistic face - unrealistic voice).

All statistical analyses were conducted with the open-source JASP software (Love et al., 2019).

3. Results

3.1. Unimodal Stimuli: Only Visual and Only Auditory Conditions
**Visual:** A 2 (Age: Old, Young) x 3 (Naturalness: Realistic, Semi-realistic, Unrealistic) mixed ANOVA on the ratings given for the visual stimuli showed a main effect of naturalness (F(2,116)=48.75, p<0.001, η²p = 0.46). Unrealistic faces elicited significantly higher uncanny scores than semi-realistic faces (t=5.30, p<0.001) and realistic faces (t=9.87, p<0.001). Also, semi-realistic faces elicited significantly higher uncanny scores than realistic faces (t=4.57, p<0.001).

There was also an interaction between age and naturalness (F(2,116)=3.88, p=0.023, η²=0.063). In the younger group, the unrealistic face elicited significantly higher uncanny scores than the realistic face (t=4.71, p<0.001) whereas in the older group there was no significant difference between the two (t=2.78, p=0.095). Similarly, in the younger group, the semi-realistic face elicited significantly higher uncanny scores than the realistic face (t=4.24, p<0.001) whereas no such difference was found between the two in the older group (t=2.23, p =0.42).

There was no effect of age on the uncanny ratings (F(1,58)=1.19, p=0.28). Bonferroni corrections were done for all multiple comparisons (Figure 3a).

**Auditory:** A 2 (Age: Old, Young) x 3 (Naturalness: Real, Semi-realistic, Unrealistic) mixed ANOVA on the ratings given for the auditory stimuli also showed a main effect of naturalness (F(2,116)=49.42, p<0.001, η²p=0.46). Unrealistic voices elicited significantly higher uncanny scores than semi-realistic voices (t=-3.10, p=0.007) and realistic voices (t=-9.73, p<0.001). Also, semi-realistic voices elicited significantly higher uncanny scores than realistic voices (t=-6.63, p<0.001).

There was also an interaction between age and naturalness (F(2,116)=4.12, p=0.019, η²p=0.066). The semi-realistic voice elicited significantly higher uncanny scores than the realistic voice in the younger group (t=-6.53, p<0.001) but not in the older group (t=-2.85, p=0.079). In addition, the unrealistic voice elicited significantly higher uncanny scores than the semi-realistic voice in the older group (t=-3.85, p=0.003) but not in the younger group (t=-0.53, p=1.0).

There was no effect of age on the uncanny ratings (F(1,58)=2.83, p=0.098). Bonferroni corrections were done for all multiple comparisons (Figure 3b).

![Figure 3](image-url) **Figure 3.** Average uncanny scores of the unimodal stimuli across age groups. (a) Visual stimuli, (b) Auditory stimuli
3.2. Multimodal Stimuli: Audio-visual conditions

A 2 (Age: Old, Young) x 3 (Visual modality: Realistic, Semi-realistic, Unrealistic) x 3 (Auditory modality: Realistic, Semi-realistic, Unrealistic) mixed ANOVA on the ratings of the audio-visual stimuli showed a main effect of visual modality (F(2,116)=71.91, p<0.001, η2p=0.45) and a main effect of auditory modality (F(2,116)=47.86, p<0.001, η2p=0.55).

Multimodal stimuli that featured unrealistic faces elicited higher uncanny scores than the ones that featured semi-realistic faces (t=-6.634, p<0.001) and realistic faces (t=-11.97, p<0.001). Also, multimodal stimuli that featured semi-realistic faces elicited higher uncanny scores than the ones that featured realistic faces (t=-5.34, p<0.001).

Multimodal stimuli that featured unrealistic voices elicited higher uncanny scores than the ones that featured semi-realistic voices (t=-4.61, p<0.001) and realistic voices (t=-9.78, p<0.001). Also, multimodal stimuli that featured semi-realistic voices elicited higher uncanny scores than the ones that featured realistic voices (t=-5.17, p<0.001).

There was also an interaction between age and visual modality (F(2,116)=3.77, p=0.026, η2p=0.061), age and auditory modality (F(2,116)=3.19, p=0.045, η2p=0.052), and visual modality and auditory modality (F(4,232)=23.38, p<0.001, η2p=0.287).

Both in the younger and the older groups, visual stimuli that featured unrealistic faces elicited higher uncanny scores than the visual featured semi-realistic faces (young:(t=-6.21, p<0.001), old:(t=-2.68, p=0.125)) and realistic faces (young:(t=-10.27, p<0.001), old:(t=-4.05, p=0.001)). Also, visual stimuli that featured semi-realistic faces elicited higher uncanny scores than the realistic faces (young:(t=-5.04, p=0.001), old:(t=-3.49, p=0.01)) in both age groups.

In the both age groups, auditory stimuli that featured unrealistic voices elicited significantly higher uncanny scores than the auditory featured realistic voices (young:(t=-8.09, p<0.001), old:(t=-5.73, p<0.001)). Also, auditory featured semi-realistic voices elicited significantly higher uncanny scores than realistic voices (t=-5.4, p<0.001) in the younger group but not in the older group (t=-1.9, p=0.89). In the older group, auditory stimuli that featured unrealistic voices elicited significantly higher uncanny scores than the auditory featured semi-realistic voices (t=-3.83, p=0.003) but there wasn’t significant difference with the same comparison in the younger group (t=-2.68, p=0.125).

Multimodal unrealistic face - unrealistic voice stimuli elicited significantly higher uncanny scores than most stimuli (p=0.038 with semi-realistic face - unrealistic voice stimuli, p=0.001 with unrealistic face - realistic voice stimuli, others p<0.001). However, unrealistic face - semi-realistic voice stimuli did not have a significant difference with unrealistic face - unrealistic voice stimuli (t=-1.68, p=1). Realistic face - realistic voice stimuli elicited significantly lower uncanny scores than other multimodal stimuli (all p<0.001).

Bonferroni corrections were done for all multiple comparisons (Figure 4). There was no effect of age (F(1,58)=0.21, p=0.646).
3.3. Congruency Between Age Group Comparisons

A 2 (Age: Old, Young) x 2 (Congruency: Congruence, Incongruence) mixed ANOVA on the ratings of the congruency situation of audio-visual stimuli pairs showed a strong main effect of congruency (F(1,29)=27.01, p<0.001, η²p=0.48). There was also an interaction between age and congruency (F(1,29)=5.53, p=0.026, η²p=0.16).

For the younger group, incongruent stimuli pairs elicited significantly higher uncanny scores than congruent stimuli pairs (t=-5.58, p<0.001). At the older group, incongruent stimuli pairs elicited significantly higher uncanny scores than congruent stimuli pairs (t=-2.8, p=0.042). However, the difference between congruent and incongruent stimuli were more pronounced in the younger group compared to the older group. Bonferroni corrections were done for all multiple comparisons (Figure 5).

3.4. Unimodal and Multimodal Stimuli Together

An omnibus 2 (Age: Old, Young) x 3 (Modality: Only Visual, Only Auditory, Audio-visual) x 3 (Naturalness: Realistic, Semi-realistic, Unrealistic) mixed ANOVA showed a main effect of modality (F(2,116)=25.73, p<0.001, η²p=0.31). Multimodal stimuli elicited higher uncanny scores than the unimodal stimuli (for visual: t=-6.30,
There was also a main effect of naturalness ($F(2,116)=110.37$, $p<0.001$, $\eta^2_{p}=0.66$). Unrealistic stimuli elicited higher uncanny scores than the semi-realistic ($t=-6.92$, $p<0.001$) and realistic stimuli ($t=-14.85$, $p<0.001$). In addition, semi-realistic stimuli elicited higher uncanny scores than the realistic stimuli ($t=-7.93$, $p<0.001$). There was no effect of age ($F(1,58)=1.22 \times 10^{-7}$, $p=1.0$).

There was also an interaction between age and naturalness ($F(2,116)=4.02$, $p=0.021$, $\eta^2_{p}=0.065$), age and modality ($F(2,116)=4.66$, $p=0.011$, $\eta^2_{p}=0.074$), and age, modality, and naturalness ($F(4,232)=3.90$, $p=0.004$, $\eta^2_{p}=0.043$). The interaction between modality and naturalness was marginally significant ($F(4,232)=2.58$, $p=0.051$, $\eta^2_{p}=0.043$). Multimodal stimuli elicited higher uncanny scores than the auditory stimuli for the realistic and unrealistic conditions (all $p<0.001$) but not for the semi-realistic condition ($t=-3.17$, $p=0.061$). In addition, multimodal stimuli elicited higher scores than the auditory stimuli for the older group but not for the younger group ($t=-2.81$, $p=0.087$), (Figure 6).

Figure 6. Average uncanny scores across all modality types and age groups.

3.5. Comparison with the control stimulus: Average uncanny scores for multimodal stimuli

Our hypothesis was that the control stimulus of the experiment, i.e. “Real face-realistic voice” would be scored as the least uncanny among all the multimodal stimuli. Indeed, the older group scored the control stimulus as the least uncanny ($M=3.25$, S.E.=0.36). However, the younger group scored the “Realistic face-realistic voice” stimuli as the least uncanny ($M=2.4$, S.E.=0.29). This is also the only stimulus that showed a significant difference between the age groups ($F(1,58)=7.35$, $p=0.009$), (Figure 7).
4. Discussion

We conducted a behavioural study to explore the uncanniness caused by audio-visual components of face and voice stimuli that represent different levels of human-robot attributes. Previous research shows that humans reportedly have more positive feelings and trust in humanlike agents than mechanical robotic agents and prefer humanlike faces and voices for interaction (Mathur & Reichling, 2016; Strait et al., 2015). We hypothesized that dissonant audio and visual stimuli pairs could violate our conceptual expectation of humans and different types of robots. Human faces and voices were evaluated in three different naturalism levels. Custom made unrealistic, semi-realistic and realistic face and voice designs were used as experimental stimuli. We investigated how multimodal stimuli and the congruence of visual and auditory aspects of the stimuli contribute to the uncanniness perception and differ from the audio and visual components.

Robotic featured face and voice designs in our experiment cause a high level of eeriness/uncanny ratings compared to the humanlike designs. As expected, the control stimulus (real face - realistic voice) is rated with normal and lower uncanniness feelings. Multimodal realistic face - voice paired stimuli are scored with the lowest uncanniness and unrealistic face - voice paired stimuli are scored with the highest uncanniness. These polarized results meet our expectation that people tend to have normal and positive feelings towards humanistic figures than unnatural mechanical figures, which can be considered as a validation of our stimulus. There is a trend of interaction between voice and face aspects, which could be a potential support for how different audios are evaluated particularly for different faces and the other way around. There is a linear trend from normal scores to uncanny scores between realistic, semi-realistic and unrealistic audio-visual stimuli. Realistic voice based

Figure 7. Average Uncanny Scores of Multimodal Stimuli.
multimodal stimuli have lower uncanniness than semi-realistic and unrealistic voice-based stimuli. We observe a bigger score gap between the realistic voice paired and the semi-realistic voice paired stimuli, compared to the semi-realistic voice paired and the unrealistic voice paired stimuli. These results show that realistic voice stimuli have a dominant role on normal feelings in audio-visual stimuli comparison.

Unimodal visual stimuli have a significant effect on uncanniness score especially in young participants. Unrealistic faces elicit higher uncanniness than semi-realistic faces and semi-realistic faces elicit higher uncanniness than realistic faces significantly in the young age group. However older participants do not react significantly to realistic and unrealistic faces. Young people's high exposure to graphical human and robot designs and familiarity may have caused layered responses to different realism levels of visual stimuli (Björling et al., 2021). Uncanniness score differences between realistic and unrealistic visual designs are less in the older group. Older people's preference over realistic faces may have resulted with higher uncanniness and high acceptance threshold to graphical designs of humans, so elderly had a smoother reaction to different realism levels of visual stimuli. Unimodal auditory stimuli have distinct responses from older and younger age groups. Younger group react with a significantly higher uncanniness score to semi-realistic faces than realistic faces; but the older group react with a significantly higher uncanniness score to unrealistic faces than semi-realistic faces. These results reveal younger people's distinctive reactions on realistic voice stimuli which occurred similarly on realistic face stimuli.

Multimodal features of both auditory and visual stimuli have a significant main effect on the uncanniness score. Visual features in multimodal stimuli, lead to significantly different reactions with all naturalness levels in both age groups, in which unrealistic faces elicited higher uncanniness than semi-realistic faces and semi-realistic faces elicited higher uncanniness than realistic faces. However, auditory features in multimodal stimuli significantly elicited higher uncanniness scores on unrealistic voices than realistic voices in both age groups, other semi-realistic and unrealistic comparisons did not have consistent and levelled responses in neither older nor in younger group. The layered pattern of responses thus is better controlled by the visual features and this could indicate a visual bias on uncanniness scores in the presence of multimodal stimuli.

Unimodal auditory, unimodal visual and multimodal audio-visual stimuli scores on uncanniness were analysed to answer the main research question as how modality affects uncanny valley. Only face and only voice conditions as unimodal stimuli scored with lower uncanniness than face - voice paired multimodal stimuli. Both unimodal realistic faces and unimodal realistic voice stimuli have lower uncanniness scores than multimodal audio-visual realistic stimuli in all age groups. Unrealistic face - unrealistic voice multimodal stimuli have the highest uncanniness score but unimodal unrealistic face and unimodal unrealistic voice stimuli have both lower uncanniness scores than multimodal unrealistic face - voice stimuli. Based on the results above, we assume that multimodal audio-visual stimuli can generate stronger uncanny feelings than
unimodal stimuli. Multimodal conditions, in which facial and speech components coexist, gave rise to the uncanny valley effects (Tinwell et al., 2015).

We obtained promising results from generation-based comparisons in our experiment. Both the older and younger age groups scored realistic face paired stimuli with the lowest uncanniness, but significant differences were found in multimodal and unimodal realistic stimuli between generations. Younger age group scored multimodal realistic face - realistic voice stimuli with the lowest uncanniness, but the older age group scored control stimulus (real face - realistic voice) with the lowest uncanniness. In addition, the younger age group scored multimodal realistic face - realistic voice stimuli and unimodal only realistic face stimuli significantly lower uncanniness than the older age group. It can be assumed that the younger generation is less sensitive and more familiar with animated and design faces of humans, but the older generation strongly prefers real human conditions rather than designs. The younger generation is exposed to a lot of graphic human design from video games and animated movies from their childhood, but graphic human designs is a relatively new concept for older people (Rose & Björling, 2017). Beyond this; older age group reacted to unimodal and multimodal stimuli with linear uncanniness from realistic to unrealistic. But the younger age group broke the linear trend and scored realistic, semi-realistic and unrealistic stimuli separately, and modality had effects on scores. Younger population can interact with robots responsively and has layered perception to robots, but older people have tended to generalize robot concepts and neglect robotic attitudes (Robb et al., 2020). This assumption can be supported with elder’s inclusive preference on humanlike robots, regardless of their function (Tu et al., 2020).

One of the cognitive explanations for the uncanny valley is violation of expectations when encountered with almost-human agents (MacDorman & Ishiguro, 2006; Saygin et al., 2012; Urgen et al., 2018). It has been thought that humans tend to attribute humanistic characteristics to human-like agents, but when they realize that they possess certain characteristics that are not human-like, they may find them uncanny. In our study, both older and younger participants rated the real human stimuli less eerie than realistic face designs. Thus, people’s discomfort with realistic designs may be linked to violation of their human concept. More specifically, face - voice mismatch effect may subvert categorical boundaries about robot and human definitions in our perception, as previous studies reported (Meah & Moore, 2014; W. J. Mitchell et al., 2011). Future work could investigate whether increasing familiarity with realistic designs would ease the discomfort people have.

Older participants have distinct uncanniness scores on realistic face - unrealistic voice stimulus than realistic face - voice stimulus. Negative feelings on incongruence stimuli suggest that mismatch between face - voice designs can cause uncanny emotions in older generations. In previous studies, older people report negative and uncanny feelings during the human-robot interaction processes (Trovato et al., 2015). Our results suggest that the conceptual congruence could be the main effect for the uncanny valley in the older population. Regarding this result, it is also important to note that different exposures of young and old populations to technological developments may cause them to develop different concepts about
robot perception and human-robot interaction.

To conclude, we find that multimodal stimuli induce distinct uncanny responses than unimodal only audio or only visual stimuli. In other words, our work extends previous work on uncanny valley which largely focused on the visual modality, and shows that auditory modality is a determinant factor for uncanny valley. In addition, multimodal stimuli that have incongruent audio and visual components elicited significantly higher uncanny scores than stimuli that have congruent components. However, the difference between congruent and incongruent stimuli were more pronounced in the younger group compared to the older group. Furthermore, our results also suggest some age differences in the uncanny valley, in particular young adults being more sensitive to the visual layering of naturalness. Future work should investigate the neural mechanisms underlying the uncanny valley effect elicited by audio-visual stimuli with electrophysiological and neuroimaging methods.

5. References


mismatch as one road to the valley of eeriness. *Frontiers in Psychology.*
https://doi.org/10.3389/fpsyg.2015.00390


