Are Virtual Humans Uncanny?: Varying Speech, Appearance and Motion to Better Understand the Acceptability of Synthetic Humans

Takaaki Kuratate, Kathryn Ayers, Jeesun Kim, Marcia Riley, Denis Burnham

MARCS Auditory Laboratories, University of Western Sydney, Australia
t.kuratate@uws.edu.au, 98159215@student.uws.edu.au, j.kim@uws.edu.au, mjriley7@gmail.com, d.burnham@uws.edu.au

Abstract

In this experiment we examine the uncanny valley effect as a guide, albeit incomplete, to designing engaging human-like interfaces. In this effect falling just short of perfection in creating synthetic humans elicits a pronounced negative reaction and here this is explored with talking head animations focusing on naturalness in speech, appearance and face motion. We implemented a set of techniques to manipulate naturalness for each of these aspects, and present the results of our study evaluating these stimuli. Outcomes of this method provide insights for the choice of the degree of realism or naturalness required for various talking heads.

Index Terms: uncanny valley, multi-modal perception, face animation

1 Introduction

Investigations into the phenomenon dubbed the “uncanny valley” by Mori in 1970 [1] have recently had something of a resurgence. Particularly driving this resurgence is commercial interest in the success of video games, movies, and other entertainment applications such as virtual worlds, and the proliferation of hand-held devices with sophisticated graphics capabilities, such as Apple’s iPhone and recent iPod models. Our focus is in exploring the mechanisms behind the acceptance or rejection of virtual talking-head animations in an effort to create engaging animations for a variety of applications, with particular interest in the animations’ auditory-visual aspects. While we rely on a variety of realism in our faces to test our ideas, we do know that realistic facial animation remains a major challenge in computer graphics research. Our models are not realistic enough to be confused with actual people, and we do not expect to replicate a theoretical uncanny valley effect. Rather, our aim is to uncover insights into designing perceptibly acceptable human representations which carry critical components of human-human communication, including speech-related visual information, into the human-virtual realm without evoking negative reactions.

These negative reactions to realistic computer-generated characters appear linked to the uncanny valley effect [2, 3]. Mori’s graph (see Figure 1) shows the uncanny valley as a place where we lose a sense of familiarity with a human-like representation. Mori explains that in some way, our expectations of authenticity no longer match our observations, and thus the familiar becomes the strange, and is rejected. Anecdotal evidence [4] of the public’s reactions points to a similar rejection of virtual humans in ‘photo-realistic’ films such as ‘Final Fantasy: The Spirits Within’(2001) and more recently ”The Polar Express”(2004). In fact, the contrast between the appeal of characters in ”The Polar Express” and ”The Incredibles”, released in the same year, show that deliberate design choices concerning realism may strongly influence box office success [5, 6].

Mori’s ideas were put forth as hypotheses in reaction to the appearance and motion of robots not yet built, and despite some anecdotal evidence in support of these ideas from the closely related field of computer animation like those noted above, empirical studies are required to elucidate the factors promoting acceptance of realistic human representations. Mori introduced the visual aspects of appearance and motion as important perceptual modalities. Other modalities such as audition may play a role, and, in addition, the contribution of interactions among modalities begs further investigation [7, 8]. Insights will potentially benefit not only robot and virtual human design for a variety of applications, but fields such as prosthetics [9] as well.

2 Experimental Method

This study manipulates speech, appearance and motion, reflecting our interest in uncovering the role individual pieces of multi-modal presentations play in causing repulsion or acceptability in synthetic human representations. These factors are combined into different talking head (TH) animations as described below, and used as stimuli along with audio-only and static conditions.
2.1 Varying Speech

Three types of speech stimuli were prepared: natural recorded human speech, sine-wave speech (SWS), and completely synthetic speech obtained from a text-to-speech (TTS) system. The sine-wave speech [10] was synthesized from the natural speech, and, in the third case, the natural speech was used to impose timing information for the TTS-based speech. (We used Apple’s TTS engine on Mac OS X, and the Xcode Developers software called “Repeat After Me” which can impose pitch and duration on TTS parameters based on recorded speech.) Note that the temporal aspects of speech were not altered, and all speech sets used the same visual cues (except, of course, the audio-only condition).

By varying speech data, we aim to determine what role, if any, speech plays in the acceptability of virtual talking heads. Our result may support the consistency theory presented in [11] where natural speech would be liked most when paired with more natural-looking faces and motion. However, since we readily accept realistic voices with cartoonish embodiments, this prediction for speech is uncertain.

2.2 Varying Appearance

Face naturalness was manipulated by morphing from a fictional character to the 3D scanned face (using a Cyberware 3030RGB scanner) of an actual person at intervals of 10%, producing 11 faces ranging from 0% to 100% in terms of their reality. Examples are shown in Figure 2. In manipulating appearance, we are looking for a response similar to that found by MacDorman and Ishiguro [8] who demonstrated an uncanny valley effect for each of two sets of static images morphed from a robot to an android, and from the android to the person serving as android model. The images rated most eerie were also those rated less familiar, and occurred at a morph between the robot and the android. Hanson’s experiment [12] shows us that by making careful design choices we can avoid the uncanny valley for static images. However, introducing additional dimensions (such as motion and speech) and their possible interactions soon renders the problem of hand tuning for acceptability intractable [13].

Figure 2: Morph from a fictional to non-fictional character. Left: Fictional; Centre: 50% fictional-50% real morph; Right: Scan of real subject.

2.3 Varying Motion

Human face motion was collected along with corresponding natural speech data from human subjects using an OPTOTRAK 3020 motion capture system (Northern Digital Inc.). We used an animation system in which face motion is represented by a small set of parameters consisting of 6 degree-of-freedom head motion parameters (rotation and translation) and linear combination values of principal components (PCs) of the face model. These parameters are then mapped onto any number of selected face models [14] of varying degrees of realism. (Gaze information is not included.) For changing face motion, then, it sufficed to alter the original PC trajectories representing the motion capture data to obtain degraded motion. Both the head and face motion were altered by applying an algorithm which skips over some of the high frequency information in the original sampling, substituting the modified trajectory for the original. (Specifically, the algorithm, implemented in MATLAB, skips over trajectory extrema when the distance between adjacent extrema was less than some percentage of the maximum adjacent distance found in a trajectory.) Original and altered motion is shown in Figure 3. When animated and played with speech data, the degraded motion provided a mismatch between the voice and face motion data reminiscent of dubbing one language for another in films.

In manipulating motion, we expected to find that natural motion would be more likable than degraded motion, a premise supported by [15], who used different body models with no motion, natural (motion captured) motion and degraded motion to study the effect of motion on acceptability. In [15], the authors found an uncanny valley for static data; and that natural motion improved acceptability except in the realistic representations, whereas degraded motion made the models less acceptable than their static counterparts.

2.4 Running the experiment

Animations were created for the conditions shown in Table 1 along with 3 audio-only conditions. Specifically, we prepared 99 (4 speech x 11 faces x 2 motions + 11 stills) movies. Note that still images were presented only with the no audio case. Motion stimuli randomly varied in sentence content over 10 sentences, so that a good degree of variability was provided over the 88 stimuli (excluding still faces, but including no-audio). We presented these stimuli in 4 blocks all in pseudo-randomized order, taking about 15 minutes per block to view and evaluate. 16 people were recruited for the study, consisting of 7 males and 9 females, with ages ranging from 18-43 years. Using the experimental setup shown in Figure 4, subjects evaluated the animations and audio-only cases for likeability and naturalness. All subjects were volunteers, spoke English as a first language, and could stop at any time for any reason during the experiment.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speech</td>
<td>Normal, Sine wave, TTS, No Sound</td>
</tr>
<tr>
<td>Face model</td>
<td>[0 - 100%] real, increment by 10%</td>
</tr>
<tr>
<td>Face motion</td>
<td>Motion capture, Degraded, Still</td>
</tr>
</tbody>
</table>

3 Results

Figure 5 shows the averaged responses of 16 subjects: each condition indicates (a) original speech, (b) still, no audio, (c) sinewave speech, and (d) TTS audio (Apple TTS “Fred”). Both naturalness and likability responses are plotted with respect to the morph scale (appearance) showing the amount of realism (0-100%) in the face model; original motion and modified motion results are also plotted. (Due to a technical glitch, the 100% real morph condition was excluded from the experiment except for the still face + no
Figure 3: An example of degraded motion (a) face motion in PC space, and (b) head motion (Z-axis rotation).

Figure 4: The experimental setup.

audio condition shown in Fig. 5(b). Thus each subject actually saw 91 animations (4 speech x10 faces x 2 motions + 11 still) and three audio-only stimuli.

From these figures, we can see the following:

- Response to extrema-based degraded motion does not significantly differ from original motion across audio conditions.
- Naturalness increases along with realism in each audio condition: i.e., as the face model approaches the 3D scan of an actual person, these responses increase. The rates of increase are almost the same across all four audio conditions.
- Likability does not show obvious changes with appearance changes as found for naturalness.
- Changing the audio condition affects the offset value (y-intercept) of both naturalness and likability.
- Overall responses are more favourable in the original speech condition: SWS and TTS speech yield lower, similar responses; however, the likability in the silent condition is better than the SWS and TTS conditions.

Figure 6 shows averaged responses from the same 16 subjects for the audio-only condition. As can be seen from this figure, the original audio fared better than the sine wave and TTS audio, which are similar in naturalness and likability ratings. (During testing, the same visual interface was used (Figure 4), but the window was left blank (black) while presenting audio stimuli.)

In summary we found that morphed faces affect naturalness, with realistic faces viewed as more natural; the original speech shows better responses in both animation conditions (original and degraded motion), and elicited better likability responses than the still, silent condition. Degraded motion did not affect responses.

3.1 Discussion and Conclusions

The original speech showed better responses over all conditions, including audio-only, than sine wave and TTS-based speech. Of interest is the result in the audio-only condition (Figure 6), which indicates that audio signals with negative responses (both sine wave speech and TTS all receive ratings lower than 4, the neutral response) may actually disturb the visual signal in both naturalness and likability dimensions. Also, as responses in the audio-only condition in Figure 6 are quite similar to the offset values (y-intercept) in Figure 5, we suspect that audio conditions are dominant in the audio-visual conditions, shifting the same general rate of change for appearance and motion either up or down.

The dominance of voice in our findings may be due to the differences among the likability of the speech stimuli, with a strong
tendency toward an almost bipolar distribution across subject responses of like versus dislike in original versus altered speech. We believe that increasing variety in the speech signals (e.g., better quality TTS voices, more robotic voices, low and high band pass filtered speech signals) and using voice morphing between signals will give us more detailed information about the role speech plays.

In light of the strong speech response, we see a need to increase the variety in our visual stimuli. When compared to the differences among the speech stimuli, the motion differences were perhaps too subtle to elicit strong perceptual responses. Similarly, our appearance models were limited to combinations of only two representations, limiting their variability. Perhaps partially because of this we saw very little change in likability as the appearance changed, and we may be testing whether these specific models were liked by the subjects, instead of testing the effects of realism in general. In fact, likability ratings did seem strongly dependent on face model preferences because subject ratings showed that responses to the original two static face models were not appreciably changed by morphing. Our next study, then, will test several different models, instead of strictly morphing between two.

Realism did show an effect on naturalness ratings, however, and in Figure 5 we see that naturalness for all cases changed as expected, increasing as realism increased.

As noted above, we did not find obvious differences in responses to the two face motion cases, and we suspect that the degraded motion was still too synchronized with the original motion at a gross level, diminishing any perceptible differences. We plan to continue exploring altered motion (i.e., unsynchronized; stronger perceptible changes) in future experiments. In addition, we are planning a speech-in-noise test to evaluate speech percep-

Figure 5: Naturalness and Likability responses by appearance (morph scale) with different audio conditions: (a) Original speech, (b) No Audio, (c) Sinewave speech and (d) TTS speech.
tibility for different representations.

Our model could be improved by including better texture information, gaze information, and internal but visible mouth structures (teeth and tongue). To this end we are acquiring more accurate face models using a high-quality 3D face scanner (3DMD scanner, http://www.3dmd.com), and will conduct further experiments using the new stimuli. However, rather than creating completely realistic computer animation, our chief aim is to find levels of realism that appear natural and engaging, and to define and identify those representations that are able to carry important cues for human perception.

By closely examining the interactions among the different modalities in talking heads we may find ways to mitigate the negative effects described by the uncanny valley. This experiment gives us a guideline and directions for future experiments in order to tease apart the different components affecting the acceptance of realistic virtual representations.

4 Acknowledgements

We acknowledge Australian Research Council (ARC) (DP0666891), and ARC and National Health and Medical Research Council Special Initiatives (TS0669874) support.

References