Game Sound Technology and Player Interaction: Concepts and Developments

Mark Grimshaw
University of Bolton, UK
Chapter 11
Uncanny Speech

Angela Tinwell
University of Bolton, UK

Mark Grimshaw
University of Bolton, UK

Andrew Williams
University of Bolton, UK

ABSTRACT
With increasing sophistication of realism for human-like characters within computer games, this chapter investigates player perception of audio-visual speech for virtual characters in relation to the Uncanny Valley. Building on the findings from both empirical studies and a literature survey, a conceptual framework for the uncanny and speech is put forward which includes qualities of speech sound, lip-sync, human-likeness of voice, and facial expression. A cross-modal mismatch for the fidelity of speech with image can increase uncanniness and as much attention should be given to speech sound qualities as aesthetic visual qualities by game developers to control how uncanny a character is perceived to be.

INTRODUCTION
As technological advancements allow for the representation of high fidelity, realistic, human-like characters within computer games, aspects of a character’s appearance and behaviour are being associated with the Uncanny Valley phenomenon. (A definition of the Uncanny Valley is provided in the first section of this chapter.) It seems that one of the main factors contributing to a character being regarded as lifeless as opposed to lifelike is the character’s speech. In 2006, Quantic Dream revealed a tech demo (The Casting) for the computer game Heavy Rain (2006), in which the main character, Mary Smith, evoked a somewhat negative responsive from the audience (Gouskos, 2006). Criticism was made of the uncanny nature of Mary Smith’s speech in that it sounded strange and out of context with the given facial expression and emotion portrayed by this character. A closer inspection of the video showed that not only were there errors in the sound recording (disparities between the acoustics and the volume and materials of the room with excessive plosives contradicting the distant camera and microphone), but a lack of correct pitch and intonation for speech and a lack...
of synchronization of speech with lip movement were factors that reduced the overall believability for this character (Tinwell & Grimshaw, 2010). A mismatch between the conveyed emotion of Mary Smith’s voice with her gestures and posture exacerbated how unnatural and odd the character was perceived to be. MacDorman (quoted in Gouskos, 2006), observed that a perceived asynchrony of lip movement with speech was one of the factors that people found disturbing about Mary Smith:

_In addition, there is sometimes a lack of synchronization with her speech and lip movements, which is very disturbing to people. People ‘hear’ with their eyes as well as their ears. By this, I mean that if you play an identical sound while looking at a person’s lips, the lip movements can cause you to hear the sound differently._

Since Mary Smith was revealed in 2006, increasing technological sophistication for computer games has allowed for heightened realism of human-like characters. Cinematic animation is achieved not only for cut scenes and trailers containing full motion video (FMV) but also for animation during in-game play. For example, the phoneme extractor and facial expression tool Faceposer designed by Valve for titles such as _Left 4 Dead_ (2008) and _Half Life 2_ (2008). However it would seem that speech, as a factor integral to the uncanny phenomenon, is often overlooked when compared to the aesthetic visual qualities of behaviour of a human-like character. So far there have been limited studies to ascertain which factors contribute to the uncanny for virtual characters. In response to the hearsay in mass media raised by characters such as Mary Smith, Tinwell and Grimshaw (2010) conducted a study to investigate how the cross-modality of image and sound might exaggerate the uncanny. The results from this study are referred to throughout all sections within this chapter as the Uncanny Modality (UM) study, unless otherwise stated from another study. Prior to this, much of the work on the uncanny had been visually-based, excluding sound as a factor. As a way towards building a conceptual framework for the uncanny and virtual characters in immersive 3D environments, this chapter defines how characteristics for a character’s speech may exaggerate the uncanny by considering aspects such as synchronization of audio and video streams, articulation, and qualities of speech.

The first section provides an exposition of the Uncanny Valley describing how the theory came about, previous investigation into the theory and potential limitations of the theory in relation to virtual characters.

Previous authors (such as Bailenson et al., 2005; Brenton, Gillies, Ballin, & Chatting, 2005; and Vinayagamoorthy, Steed, & Slater, 2005) have suggested that uncanniness is increased when the behavioural fidelity for a realistic, human-like character does not match up with that character’s realistic, human-like appearance. The second section discusses how a cross-modal mismatch between a character’s appearance and speech may exaggerate the uncanny. For instance whether a character’s speech may be perceived as belonging to a character or not, based on that character’s appearance.

The third section discusses how particular qualities of speech such as slowness of speech, intonation and pitch and how monotone the voice sounds, may influence perceived uncanniness and how such qualities might work to the advantage of those characters intended to elicit an eerie sensation.

The results from the UM study (Tinwell & Grimshaw, 2010) revealed a strong relationship between how strange a character is perceived to be and the lack of synchronization of speech and lip movement. (Characters rated as close to perfect synchronization for lip movement and speech were perceived as less strange than those with disparities in synchronization.) The fourth section reviews the findings from this study and also puts forward future experiments that may...
help to define acceptable levels of asynchrony for computer games where uncanniness is not desired.

For figures onscreen an over exaggeration of pronunciation for particular words can make the figure appear uncanny to the viewer as the figure seems absurd or comical (Spadoni, 2000). The fifth section considers how the manner of articulation of speech may influence the uncanny by examining the visual representation (viseme) for each phoneme within the choreography tool Faceposer (Valve Corporation, 2008).

A summary is presented in the final section that defines the outcomes from this inquiry as to how speech influences the uncanny for realistic, human-like virtual characters as a way towards building a conceptual framework for the uncanny. It is intended that this framework is not only relevant to computer game characters but also for characters within a wider context of user interfaces. For example virtual conversational agents within therapeutic applications used to interact with autistic children to aid the development of communication skills. Also those virtual conversational agents used to deliver learning material to students within e-learning applications.

THE UNCANNY VALLEY

The subject of the uncanny was first introduced in contemporary thought by Jentsch (1906) in an essay entitled On the Psychology of the Uncanny. Jentsch described the uncanny as a mental state where one cannot distinguish between what is real or unreal and which objects are alive or dead. In 1919, to establish what caused certain objects to be construed as frightening or uncanny, Sigmund Freud made reference to Jentch’s essay as a way to describe the feeling caused when one cannot detect if an object is animate or inanimate upon encountering objects such as “waxwork figures, ingeniously constructed dolls and automata” (p. 226). Freud characterized the uncanny as similar to the notion of a doppelganger; the body replica being at first an assurance against death, then the more sinister reminder of death’s omen “a ghastly harbinger of death” (p. 235).

Building on previous depictions of the uncanny, the roboticist Masahiro Mori (1970, as translated by MacDorman & Minato, 2005) observed that a robot continued to be perceived as more familiar and pleasing to a viewer as the robot’s appearance became more human-like. However, a more negative response was evoked from the robot as the degree of human-likeness reached a stage at which the robot was close to being human, but not fully. Mori plotted a perpendicular slope climbing as the variables for perceived human-likeness and familiarity increased until a point was reached where the robot was regarded as more strange than familiar (see Figure 1). At this point (about 80-85% human-likeness), due to subtle deviations from the human-norm and the resounding negative associations with the robot, Mori drew a valley shaped dip. A real human was placed, escaping the valley, on the other side. Mori gave examples of objects such as zombies, corpses and lifelike prosthetic hands that lie within the valley. He also predicted that the Uncanny Valley would be amplified with movement as opposed to the still images of a robot.

Mori recommended that for robot designers, it was best to avoid designing complete androids and to instead develop humanoid robots with human-like traits, aiming for the first valley peak and not the second which would risk a fall into the Uncanny Valley. As computer game designers working in particular genres continue the pursuit of realism as a way to improve player experience and immersion, designers have the second peak as a goal to achieve believably realistic, human-like characters (Ashcraft, 2008; Plantec, 2008). To reach this goal and to assess if overcoming the Uncanny Valley is an achievable feat, further investigation and analysis of the factors that may exaggerate the uncanny is required.
Previous Investigation into the Uncanny Valley

Since Mori’s original theory of the Uncanny Valley over thirty years ago, the increasing realism possible for virtual characters and androids has sparked a renewed interest in the phenomenon (Green, MacDorman, Ho, & Vasudevan, 2008; Pollick, in press; Steckenfinger & Ghazanfar, 2009). However, there have been few empirical studies conducted to support the claims of uncanny virtual characters and androids evident within new media (Bartneck, Kanda, Ishiguro, & Hagita, 2009; MacDorman and Ishiguro, 2006; Pollick, in press; Steckenfinger & Ghazanfar, 2009).

Still images of both virtual characters and robots have been used for experiments investigating the Uncanny Valley. Design guidelines have been authored to help realistic, human-like, characters escape from the valley (for example, Green et al., 2008; MacDorman, Green, Ho, & Koch, 2009; Schneider, Wang & Yang, 2007; Seyama & Nagayama, 2007). MacDorman et al. focused on how facial proportions, skin texture and how levels of detail affect the perceived eeriness, human likeness, and attractiveness of virtual characters. Schneider et al. investigated the relationship between human-like appearance and attraction with the results indicating that the safest combination for a character designer seems to be a clearly non-human appearance with the ability to emote like a human.

Hanson (2006) conducted an experiment using still images of robots across a spectrum of human-likeness. An image of a human was morphed to an android on one half of the spectrum and then the android to a mechanical-looking, humanoid robot on the other half. The results depicted an uncanny region between the mechanical-looking, humanoid robot and the android. In a second experiment, Hanson found that it was possible to remove the uncanny region within the same plot, where it had previously existed, by changing the appearance of the android’s features to a more “cartoonish” and friendly appearance.

However the results from these experiments only provide a somewhat limited interpretation of perceived uncanniness based on inert (unresponsive) still images. Most characters used in animation and computer games are not stationary, with motion, timing and facial animation being the main factors contributing to the Uncanny Valley
Uncanny Speech

(Richards, 2008; Weschler, 2002). For realistic androids, behaviour that is natural and appropriate when engaging with humans, referred to as “contingent interaction” by Ho, MacDorman, and Pramono (2008, p. 170), is a key factor in assessing a human’s response to an android (Bartneck et al., 2009; Kanda, Hirano, Eaton, & Ishiguro, 2004). Previous authors (such as Green et al., 2008; Hanson, 2006; MacDorman et al., 2009; Schneider et al., 2007) state that the conclusions drawn from their experiments where still images had been used may have been different had movement (and sound) been included as a factor.

The perception of the uncanny does not always have to provide a negative impact for the viewer (MacDorman, 2006). The principals of the uncanny theory can work to the advantage of engineers when designing robots with the purpose of being unnerving within an appropriate setting and context. Similarly, the uncanny may help in the success of the horror game genre for zombie-type characters. Building on these findings, Tinwell and Grimshaw (2010) conducted the UM study, using video clips with sound, to investigate how the uncanny might enhance the fear factor for horror games. The results showed that combined factors such as appearance and sound can work together to exaggerate the uncanny for virtual characters.

Not only was it suggested that a lack of lip/vocalization synchronization reduced how familiar a character was perceived to be, but a perceived lack of human-likeness for a character’s voice, facial expression, and doubt in judgement as to whether the voice actually belonged to the character or not, also reduced perceived familiarity.

Limitations of Mori’s Theory

Recent studies demonstrate weaknesses within the Uncanny Valley theory and suggest it may be more complex than the simplistic valley shape that Mori plotted in his original diagram (see Figure 1). Various factors (including speech) can influence how uncanny an object is perceived to be (Bartneck, et al., 2009; Ho et al., 2004; Minato, Shimda, Ishiguro, & Itakura, 2004; Tinwell & Grimshaw, 2009). Attempts to plot Mori’s Uncanny Valley shape cannot confirm the two-dimensional construct that Mori envisaged. The results from experiments that have been conducted using cross-modal factors such as motion and sound imply that it is unlikely that the uncanny phenomena can be reduced to the two factors, perceived familiarity and human-likeness, and is instead a multi-dimensional model (see Figure 2).

Figure 2. The Uncanny Wall, (Tinwell & Grimshaw, 2009)
When ratings for perceived familiarity were plotted against human-likeness, the results from Tinwell and Grimshaw’s experiment, using 100 participants and 15 videos ranging from humanoid to human with character vocalization, depict more than one valley shape. The plot is more complex than Mori’s smooth curve and the valley shapes less steep than Mori’s perpendicular climb. The most significant valley occurs between the humanoid character Mario, on the left and the stylized, human-like Lara Croft, on the right. The nadir for this valley shape is positioned at about 50-55% human-likeness that is lower than Mori’s original prediction of 80-85% human-likeness.

Results from studies using robots with motion and speech are also inconsistent with Mori’s Uncanny Valley. MacDorman (2006) plotted ratings for perceived familiarity against human-likeness for an experiment using videos of robots from mechanical to human-like, including some stimuli with speech. The results showed no significant valley shape in keeping with the depth and gradient of Mori’s plot and that robots rated with the same degree of human-likeness can have a different rating for familiarity. Bartnek et al. (2009) found that when a robotic copy of a human was compared to that human for the two conditions movement (with motion and speech) and still, despite a significant difference in perceived human-likeness between the human and the android, there was no significant difference between perceived likeability for the android and the human. These results imply that movement may not be the only factor to influence the uncanny. Further investigation is required to assess how speech may contribute to a more multi-dimensional model to measure the uncanny.

Uncertainty exists as to whether the meaning for Mori’s original concept may have been “lost in translation” (Bartnek et al., 2009, p. 270). The word that Mori used in the title for the Uncanny Valley is *bukimi*, which, translated in Japanese, stands for “weird, ominous, or eerie”. In English, “synonyms of uncanny include unfamiliar, eerie, strange, bizarre, abnormal, alien, creepy, spine tingling, inducing goose bumps, freakish, ghastly and horrible” (MacDorman & Ishiguro, 2006, p. 312) while Freud used the word *unheimlich* to define the uncanny: Further confusing the issue, the root *heimlich* has two meanings viz familiar or agreeable and that which is concealed and should be kept from sight. Freud discussed both meanings in his 1919 essay and they are not necessarily mutually exclusive as we show below. However, despite a generic understanding for the word that Mori used, the appropriateness of the term *shinwa-kan*, (translated as *familiarity*) that Mori used in his original paper as a variable to measure and describe uncanniness has been addressed by previous authors.

As an uncommon word within Japanese culture there is no direct English equivalent for the word *shinwa-kan*. The word familiarity stands for the opposite to unfamiliarity (one of the synonyms for *bukimi*), yet the word familiarity may be open to misinterpretation. Whilst strange is a typical term for describing the unfamiliar, familiarity might be interpreted with a variety of meanings including how well-known an object appears: for example, a well-known character in popular culture or an android replica of a famous person. Bartnek et al. (2009) proposed that with no direct translation *shinwa-kan* could be treated as a “technical term” in its own right however this may cause problems when comparing the results from one experiment to another where the more generic translation “familiarity” is used as the dependent variable (p. 271). Other words such as likeability (Bartnek et al., 2009) or *unstrange* (the opposite to strange) may be closer to Mori’s original intention, nevertheless the validity for experiments conducted into the uncanny may be more robust if a standard word were to be used as a dependent variable to measure and describe perceived uncanniness: that word has yet to be agreed upon.

Conflicting views exist as to whether it is actually possible to overcome the Uncanny Valley. One theory put forward is that objects may appear
less uncanny over time as one grows used to a particular object. Brenton et al. (2005) give the example of the life-like sculpture *The Jogger* by Duane Hanson: The sculpture will appear “less uncanny the second time that it is viewed because you are expecting it and have pre-classified it as a dead object”. The effect of habituation may also apply to those with regular exposure to realistic human-like virtual characters. 3D modellers working with this type of character or gamers with an advanced level of gaming experience may be less able to detect flaws within a particular character because they had grown accustomed to the appearance and behaviour for that character by interacting with it on a regular basis (Brenton, et al., 2005). Recent empirical evidence goes against this theory. The results from a study by Tinwell and Grimshaw (2009) showed that the level of experience for both playing computer games and of using 3D modelling software made little difference in detecting uncanniness. (Judgements for those with an advanced level of experience for perceived familiarity and human-likeness had no significant difference between those with lesser or no experience.)

Tinwell and Grimshaw suggest it may never be possible to overcome the Uncanny Valley as a viewer’s discernment for detecting subtle nuances from the human norm keeps pace with developments in technology for creating realism. With a lack of empirical evidence to support the notion of an Uncanny Valley, the notion of an Uncanny Wall may be more appropriate (see Figure 2). Viewers who may at first have been “wowed” by the apparent realism of characters such as Quantic Dream’s Mary Smith (2006) or characters in animation such as *Beowulf* (Zemeckis, 2007) or *The Polar Express* (Zemeckis, 2004), soon developed the skills to detect discrepancies for such characters’ appearance and behaviour. Indeed, as soon as the next technological breakthrough in achieving realism is released, a viewer may be reminded of the flaws for a character that at first did not seem uncanny. In addition to the meaning of uncanny as used in the Uncanny Wall hypothesis being an exposition of the first Freudian sense of *heimlich/unheimlich* as described above, the undesired unmasking of the technological processes used in the production of a character, and the perception of those processes as flaws in the presentation of that character, allows us simultaneously and without contradiction to use the second meaning of *heimlich*: that which should remain out of sight. The concept of the Uncanny Wall (as opposed to the Uncanny Valley which always holds out the hope for a successful traversal to the far side), evokes a variety of myths, legends and modern stories (Frankenstein’s monster, for example, or the Golem) in which beings created by man are condemned to forever remain pale shades of those created by gods.

Further studies would be required to provide evidence for the Uncanny Wall to substantiate the hypothesis that the Uncanny Valley is an impossible surmount for realistic, human-like virtual characters. As soon as the next character is released, announced as having overcome the Uncanny Valley, we intend to conduct another test using the same characters as in the previous experiment. If those characters previously rated as close to escaping the valley, such as Emily (Image Metrics, 2008), are placed beneath the new character as perceived strangeness increases, our prediction may be justified. In the meantime, a conceptual guide for uncanny motion and sound in virtual characters may be beneficial in aiding computer game developers to manipulate the degree of uncanniness.

**CROSS-MODAL MISMATCH**

For androids, if a human-like appearance causes us to evaluate an android’s behaviour from a human standard, we are more likely to be aware of disparities from human norms (MacDorman & Ishiguro, 2006; Matsui, Minato, MacDorman, & Ishiguro, H., 2005; Minato et al., 2004). Ho et
al. (2008) observed that a robot is eeriest when a human-like appearance creates an expectation of a human form when non human-like elements fail to deliver to expectations. Also, a mismatch in the human-likeness of different features for a robot, for example, a nonhuman-like skin texture combined with human-like hair and teeth, elicited an uncanny sensation for the viewer.

With regards to virtual characters it has been suggested that a high graphical fidelity for realistic human-like characters raises expectations for the character’s behavioural fidelity (Bailenson et al., 2005; Brenton et al., 2005; Vinayagamoorthy et al., 2005). Any discrepancies from the human-norm with how a character spoke or moved would appear odd. For humanoid or anthropomorphic characters with a lower fidelity of human-likeness (for example, Mario or Sonic the Hedgehog), differences from the human-norm would be more acceptable to the viewer: Expectations are lowered based on the more stylized and iconic appearance for that character. Despite seemingly strange behaviour with jerky movements or a less than human-like voice, the viewer will still develop a positive affinity with the character. Empirical evidence implies that humanoid and anthropomorphic type characters do escape the valley dip as Mori predicted, being placed before the first peak in the valley (Tinwell, 2009; Tinwell & Grimshaw, 2009).

Evidence shows that for virtual characters (and robots) a perceived mismatch in the human-likeness for a character’s voice based on that character’s appearance exaggerates the uncanny. As part of the Uncanny Modality survey (Tinwell & Grimshaw, 2010), 100 participants rated how human-like the character’s voice sounded and how human-like the facial expression appeared using a scale from 1 (nonhuman-like) to 9 (very human-like). Strong relationships were identified between the uncanny and perceived human-likeness for a character’s voice and facial expression. The less human-like the voice sounded, the more strange the character was regarded to be. Uncanniness also increased for a character the less human-like the facial expression appeared.

Laurel (1993) suggests that to achieve harmony, there is an expectation for the sensory modalities of image and sound to have the same resolution. So that there is accord between visual appearance and behaviour for virtual characters we put forward that the degree of fidelity of human-likeness for a character’s voice should match that character’s appearance, or otherwise risk discord for that character. To avoid the uncanny, attention should be given to the fidelity of human-likeness for a character’s voice in accordance with that character’s appearance. For high fidelity human-like characters it is expected that that character should have a human-like voice of a resolution that matches their realistic, human-like appearance. However for mechanical-looking robots, a less human-like and more mechanical-sounding voice is preferable. The humanoid robot Robovie was intentionally given a mechanical sounding voice so that it appeared more natural to the viewer (Kanda et al., 2004). A voice that was too human-like may have been regarded as unnatural based on the robot’s appearance, thus exaggerating the uncanny for the robot.

To test the Uncanny Valley theory with virtual characters, it has been suggested that it is not necessary to include characters from computer games as the level of realism achieved from gaming environments generated in real-time is less than that achieved for animation and film (Brenton et al., 2005). Some characters created for television and film have been proclaimed as overcoming the Uncanny Valley: In 2008, Plantec hailed the character Emily as finally having done so. Walker, of Image Metrics, states that whilst computer games would benefit from these more realistically rendered faces, it is not yet possible to achieve the same high level of polygon counts for in-game play as achieved for television and film due to technical restrictions: “We can produce Emily-quality animation for games as well, but
it just can’t work in a real-time gaming environment” (as quoted in Ashcraft, 2008).

Accordingly, for virtual characters used within computer games that are approaching levels of realism as achieved for the film industry, it may be advisable to reduce the level of human-likeness for a character’s voice to a level that is in keeping with that character’s appearance. Actors’ voices are typically used for realistic, human-like characters’ speech in computer games. Yet, if the level of fidelity for achieving human-like realism for computer games is less than that achieved for film, a less than human-like voice should be used to avoid the character being perceived as unnatural.

Hug (2011) makes a similar point when discussing the similarities between indie game and animation film aesthetics. Hug describes an affinity between sound used in animation film or cartoons matches and the aesthetic style for the animation: “[S]ounds that are more or less de-naturalized in a comical, playful, or surreal way, which is characterized by a subservive interpretation of sound-source associations”. He further uses the example of an explosion that occurs within the arcade game Grey Matter (McMillen, Refenes, & Baranowsky, 2008) as an intriguing case of “cartoonish” sound design “when an abstract dot hits a flying cartoon brain, the latter ‘explodes’ with sounds of broken glass”. Although a more cartoonish style of sound is used for the explosion, the sound seems more in keeping with the stylized appearance of the object to which the sound belongs. The visceral sounds of the impact are still evident despite the more simplistic nature of the sound. The acoustics appear more natural as the level of detail appears to match the stylized aestheticism of the film’s environment.

Of course we do not suggest that cartoon-like voices be used with characters that are approaching believable realism in computer games, however the level of human-likeness may be subtly modified so that the perceived style of the voice sound matches the aesthetic appearance of the character. This absurd juxtaposition may be necessary to reduce the uncanny for computer game characters due to the fact that they will always be playing catch up to the level of realism achieved for film. Refinements made to character’s voices over a spectrum of human-likeness ranging from human-like to mechanical, may perhaps help to remove the uncanny where it was previously evident.

Reiter notes that recently, more attention has been given to the quality of sound in computer games to keep up with the quality of realism achieved visually for in-game play and to provide a more cinematic experience. As a method of communication both diegetic and non-diegetic game sound enhances a game’s plausibility in that sound can “trigger emotions and provide additional information otherwise hard to convey” (Reiter, 2011). Distinctions made as to the quality of game sound are not simply due to the level of clarity, resolution, or digital output achievable for sound: “Perceived quality in game audio is not a question of audio quality alone” (Reiter, 2011). For speech, textures, emotive qualities and delivery style are attributes that contribute to the perceived quality and overall believability for a character. (Qualities of speech and the uncanny are discussed further in the following section.)

Quality of speech is critical in portraying the emotive context of a character convincingly. However with regards to the uncanny, if the perceived realism and quality for a voice goes beyond that of the quality and realism for a character’s appearance, such a cross-modal mismatch could exaggerate the uncanny. Further experiments are required to test this theory. Building on the premise of Hanson’s (2006) experiment where the uncanny was removed from a morphed sequence of images from robot to human by making a robot’s features more “cartoonish” and friendly, similar changes could be made to the acoustics of speech for videos of realistic, human-like characters. Whilst the videos of characters would remain constant, the speech sound would be changed across a spectrum of human-likeness from mechanical to human-like. If our predictions are correct, char-
characters will be perceived as more strange when the speech sounds too mechanical or too human-like in relation to the fidelity of human-likeness for a character’s appearance. A character may appear more natural and be perceived as more familiar once the fidelity of human-likeness for speech is adjusted to be regarded as matching that of a character’s appearance.

QUALITIES OF SPEECH

Bizarre qualities and textures of speech served to gratify the pleasure humans sought in frightening themselves with early horror film talkies, for example the monster in Browning’s (1931) film Dracula. Some cinematic theorists argue that the success of films such as Dracula was due to an uncanny modality that occurred during the transition between silent to sound cinema (Spadoni, 2000, p. 2). Sounds that may have been perceived as unreal or strange due to technical restrictions of sound recording and production at the time were used to the advantage of the character Dracula.

For early sound film, to produce the most intelligible dialogue for the viewer, the recording process required that words were pronounced slowly, emphasizing every “syl-la-ble” (Spadoni, 2000, p. 15). However, whilst words could be easily interpreted by the viewer, this impeded delivery style made the speech sound unnatural and unreal. Delivery of speech style also influenced how strange Dracula was perceived to be.

In the role of Dracula, the acoustics of Bela Lugosi’s speech set the standard for what the “voice of horror” should be (Spadoni, 2000, pp. 63-70). The weird textures of Bela Lugosi’s voice were manipulated to create a greater conceptual peculiarity for the viewer, thus setting the eponymous character apart from other horror films. The distinctive vocal tone and pronunciation of Dracula’s speech were characteristics that critics acclaimed as the most shocking and chilling; “slow painstaking voices pronouncing each syllable at a time like those of radio announcers filled the theatre” (p. 64). As Tinwell and Grimshaw state, paraphrasing Spadoni, (2010) the unique textures and delivery style for Dracula’s speech increased the uncanny for Dracula:

Dracula’s voice, the ethereal voice of the undead, is compared to the voice of reason and materiality that is Van Helsing’s. In the former, the uncanny is marked by uneven and slow pronunciation, staggered rhythm and a foreign (that is, not English) accent and all this produces a disconnect between body and speech. Van Helsing’s speech, by contrast, is the embodiment of corpiorality; authoritative, clearly enunciated and rational in its delivery and meaning.

For zombie characters in computer games, comparisons have been made with horror film talkies as to the methods used to create and modify sound to induce an ambience of fear (Brenton et al., 2005; Perron, 2004; Roux-Girard, 2011; Toprac & Abdel-Meguid, 2011). Results from the UM study by Tinwell and Grimshaw (2009) to define cross-modal influences of image and sound and the uncanny in virtual characters show that particular qualities of speech (similar to those observed for early horror talkies) can exaggerate how uncanny a virtual character is perceived to be. Thirteen video clips of one human and twelve virtual characters in different settings and engaged in different activities were presented to 100 participants. The twelve virtual characters consisted of six realistic, human-like characters: (1) the Emily Project (2008) and (2) the Warrior (2008) both by Image Metrics; (3) Mary Smith from The Casting (Quantic Dream, 2006); (4) Alex Shepherd from Silent Hill Homecoming (Konami, 2008) and two avatars (5) Louis and (6) Francis from Left 4 Dead (Valve, 2008); four zombie characters, (7) a Smoker, (8) The Infected, (9) The Tank and (10) The Witch from Left 4 Dead; (11) a stylised, human-like Chatbot character “Lillien” (Daden Ltd, 2006); (12) a realistic, human-like
Uncanny Speech

zombie (Zombie 1) from the computer game *Alone in the Dark* (Atari Interactive, Inc, 2009) and (13) a human.

Table 1 shows the median ratings for a character’s strangeness and for the speech qualities: whether the speech seemed (a) slow, (b) monotone, (c) of the wrong intonation, (d) if the speech did not appear to belong to a character, or (e) none of the above. Characters with the same median value for strangeness were grouped together and the median values for speech qualities were then calculated for those characters or groups. (Median values were used to indicate a central tendency for results, to help establish a clear overall picture of the vital relationships over multiple qualities of speech.) The results implied that, slowness of speech, an incorrect intonation, and pitch and how monotone the voice sounded increased uncanniness.

A strong indirect relationship was identified between individual ratings for the variables “the speech intonation sounds incorrect” and “the voice belongs to the character”. This implies that if the intonation for a character’s voice is in keeping with what the viewer may have expected, this characteristic may contribute to the overall believability for that character. The two zombies the Witch and the Tank, from the computer games *Left 4 Dead* (Valve, 2008), were regarded as the most uncanny with a median strangeness rating of just 2 (see Table 1). However it seems the unintelligible hisses and snarls from the Tank were regarded as sounds that this character was likely to make based on the Tank’s appearance and how he behaved. Likewise the inhuman cries and screeches from the Witch matched her seemingly pathetic and wretched appearance. Such sounds enhanced the believability of these characters as they were in keeping with their nonhuman-like appearance.

The findings from the UM study provide empirical evidence to support the claims made by MacDorman (as quoted in Gouskos, 2006) that Mary Smith’s speech was one of the main contributing factors as to why she was perceived as uncanny. Twenty percent of participants observed a lack of correct pitch and intonation for Mary Smith’s speech. This implies that the pitch and tone for her voice may not have matched the facial expression exhibited by this character. The emotive qualities of speech may have seemed either inappropriate or out of context with how this character appeared to look and behave. The facial expression may not have matched nor accurately conveyed the emotive qualities of her voice. Attributes such as these raised doubts as to whether the voice actually belonged to this character or not, thus increasing the sense of perceived eeriness for this character.

Table 1. Median ratings for speech qualities for those characters or groups with the same median strangeness value. (Tinwell & Grimshaw, 2010). Note. Judgements for strangeness were made on 9-point scales (1 = very strange, 9 = very familiar)

<table>
<thead>
<tr>
<th>Character or Group</th>
<th>Slow</th>
<th>Monotone</th>
<th>Wrong Intonation</th>
<th>Belongs</th>
<th>None</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Tank, The Witch, (Mdn = 2)</td>
<td>10</td>
<td>9.5</td>
<td>23.5</td>
<td>56.5</td>
<td>16.5</td>
</tr>
<tr>
<td>The Infected, The Smoker, Zombie 1, Chatbot, (Mdn = 3)</td>
<td>24</td>
<td>21.5</td>
<td>40</td>
<td>42</td>
<td>8.5</td>
</tr>
<tr>
<td>Mary Smith, (Mdn = 4)</td>
<td>8</td>
<td>3</td>
<td>20</td>
<td>20</td>
<td>8</td>
</tr>
<tr>
<td>The Warrior, Alex Shepherd, (Mdn = 6)</td>
<td>14</td>
<td>17</td>
<td>17</td>
<td>62.5</td>
<td>7.5</td>
</tr>
<tr>
<td>Louis, Francis, (Mdn = 7)</td>
<td>2.5</td>
<td>3.5</td>
<td>6.5</td>
<td>79.5</td>
<td>4.5</td>
</tr>
<tr>
<td>Emily, (Mdn = 8)</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>87</td>
<td>6</td>
</tr>
<tr>
<td>Human, (Mdn = 9)</td>
<td>1</td>
<td>15</td>
<td>4</td>
<td>72</td>
<td>6</td>
</tr>
</tbody>
</table>
As well as being regarded of the wrong pitch, speech that is delivered in a slow, monotone way increased the uncanny for both zombie characters and human-like characters not intended to contest a sense of the real. Within the UM study, the Chatbot character received a less than average rating for perceived familiarity and was placed with three other zombie characters with a median strangeness value of just three (see Table 1). The Chatbot’s voice was rated individually as being slow (75%), monotone (59%), and of an incorrect intonation (76%). The “speech” for Zombie 1, grouped with the Chatbot character with a median strangeness value of three, was also judged individually as being monotone (29%), slow (42%), and of an incorrect intonation (34%). Including such qualities of speech for the zombie may have been a conscious design decision by developers to increase the perceived eeriness for a character intended to elicit an uncanny sensation. (As mentioned above, such qualities enhanced the overall impact for the monster Dracula.) However the crippled speech style for the Chatbot appeared unnatural and unreal. Such qualities for this character’s speech were factors that viewers found most annoying and irritating, exaggerating the uncanny for this character when perhaps this was not intended.

Our results imply that uncanniness is increased if speech is judged to be of the wrong pitch, too monotone, or slow in delivery style. Whilst such qualities can work to the advantage of antipathetic characters by increasing the fear factor, these qualities may work against empathetic characters in the role of hero or protagonist within a game. A designer may wish the player to have a positive affiliation with the protagonist character, yet the designer may unwittingly create an uncanny sensation for the player with speech qualities that sound strange to the viewer. Speech prerecorded in a manner that is too slow or monotone to aid clarity for post-production purposes may be judged as unnatural and should be instead recorded at an appropriate tempo. Pitch and tone of speech that do not match the facial expression or given circumstance for a character may be regarded as out of context and confusing for a viewer. To avoid the uncanny, attention should be given to ensuring that the pitch of voice accurately depicts the given emotion for a character and, once speech has been recorded at the correct pitch, that the facial expression conveys that emotion convincingly.

LIP-SYNCHRONIZATION VOCALIZATION

The process of matching lip movement to speech is an integral factor in maintaining believability for an onscreen character (Atkinson, 2009). For first-person shooters (FPS) and other similar types of action game, there are limited periods during gameplay when attention is focused solely on a headshot of a speaking character. Close up shots of a player’s character, comrades or antagonists are predominantly used when exchanging information during gameplay or during cinematic cut scenes and trailers.

The music genre of computer games provides an outlet for musicians to promote and sell their work (Kendall, 2009; Ripken, 2009). As well as FPS games, music games can provide a challenge for developers with regards to facial animation and sound. The Beatles: Rock Band (EA Games, 2009) highlights the recent success of the merger of music and computer games that use realistic, human-like characters to represent music artists. It has been found, however, that uncanny traits can leave viewers dissatisfied with particular characters within the context of a computer game (Tinwell, 2009). With emphasis directed at a character’s mouth as the vocals are matched to the music tracks, it is important that an artist’s identity be transferred effectively within this new medium (Ripken, 2009). Factors such as asynchrony may result in a negative impact on the overall believability for such characters.

This section discusses the outcomes of a lack of synchrony for lip-vocalization narration in film
Uncanny Speech

and television and the corresponding implications for characters in computer games.

**Lip Syncing for Television and Film**

The process of a viewer accepting that sound and image occur simultaneously from one given source is referred to as *synchresis* (Chion, 1994) or *synchrony* (Anderson, 1996). For early sound cinema, various methods of sound recording and post production techniques were applied before a viewer no longer doubted that a voice actually belonged to a figure onscreen. A perceived lack of synchronization between image and sound has been equated with much of the uncanny sensation evoked by films within the horror genre in early sound cinema (Spadoni, 2000, pp. 58-60). Errors in synchrony evoked the uncanny for a scene in Browning’s *Dracula* (1931). As a figure’s lips remained still, human laughter resonated within the scene. With no given body or source, the laughter is regarded as an eerie, disembodied sound.

Whilst technology allows for some improvement with cinema speakers, televisions and personal computers, most sound is still delivered through some mechanism that is physically disjunct from the onscreen image (for example, via headphones or separate speakers). Tinwell and Grimshaw (2010) note that future technologies may overcome issues with asynchrony within the broadcasting industry: “Presumably, there will be no need for such perceptual deceit once flat-panel speakers with accurate point-source technology provide simultaneously a visual display” (p. 7).

For human figures in television and film, viewers are more sensitive to an asynchrony of lip movement with speech than for visual information presented with music (Vatakis & Spence, 2005). Viewers are also more sensitive to asynchrony when sound precedes video and less so when sound lags behind video (Grant et al., 2004). Grant et al. found that for continuous streams of audio-visual speech presented onscreen, detectable asynchrony occurred at 50ms when sound preceded video, with a smaller window of acceptable asynchrony for when sound lagged behind video at 220ms. Standards set by the television broadcasting industry require that the audio stream should not precede the video stream by more than 45ms and that the audio stream should not lag behind the video stream by more than 125ms (ITU-R, 1998).

An asynchrony for speech with lip movement can lead to one misinterpreting what has been said: the McGurk Effect (1976). As a viewer, one can interpret what has been heard by what has been seen. Depending on which modality one’s attention may be drawn to for audio-visual speech (and depending on which syllable is used), the pronunciation of a visual syllable can take precedence over the auditory syllable. Conversely a sound syllable can take precedence over the visual syllable. Alternatively, as one comprehends the visual articulatory process of speech both automatically and subconsciously, one can combine the sound and visual syllable information to create a new syllable. For example, a visual “ga” coinciding with the sound “ba” can be interpreted as a “da” sound. (This type of effect was observed by MacDorman (2006) for the character Mary Smith’s speech, who was criticized for being uncanny.)

A viewer’s overall enjoyment of a television programme can be disrupted if delays occur between transmission devices for video and audio signals. To prevent confusion or irritation for the viewer, sub-titles are often preferred to dubbing of speech for foreign works. (Hassanpour, 2009).

Errors in the synchronization of lip movements with voice for figures onscreen (lip sync error) can result in different responses from the viewer depending upon the context within which the errors are portrayed. A study by Reeves and Voelker (1993) found that not only is lip sync error potentially stressful for the television viewer, but it can also lead to a dislike for a particular program and viewers evaluating the people displayed on the screen more negatively and as “less interesting, more unpleasant, less influential, more agitated, more confusing, and less successful” (p. 4). On the
contrary, lip sync error has also been deliberately used to provoke a humorous affect for the viewer where the absurd is regarded as comical as opposed to annoying. For example, the intentionally bad dubbing for characters in “Chock-Socky” movies (Tinwell & Grimshaw, 2010).

**Lip Syncing for Computer Games**

With increasing technological sophistication in the creation of realism in computer games, text-based communication systems have been replaced with virtual characters using actors’ voices. To create full voice-overs for characters, automated lip-syncing tools extract phoneme sounds from prerecorded lines of speech. The visual representation (viseme) for a particular sound is retrieved from a database of predetermined mouth shapes. Muscles within the mouth area for a 3D character are modified to create a particular mouth shape for each phoneme. Interpolated motion is inserted between the next phoneme and associated mouth shape to enable contingency of lip movement for words within a given sentence. For example, a specific mouth shape can be selected for the sound “sh” to be used in conjunction with other sounds within a word or line of speech. Full voice-overs for characters were generated for titles developed by Valve such as *Left 4 Dead* (2008) and *Half Life 2* (2008) using this technique. A phoneme extractor tool within *Faceposer* allowed for the detection and extraction of phoneme sounds from prerecorded speech to be synchronized with a character’s lips.

Whilst research has been undertaken to improve the motion quality of real-time data driven approaches for realistic visual speech synthesis (Cao, Faloustsos, Kohler, & Pighin, 2004), prior to the UM study (Tinwell & Grimshaw, 2010) there have been no attempts to investigate what impact lip-synchronization may have on viewer perception and the uncanny in virtual characters. Videos of 13 virtual characters ranging from humanoid to human were rated by 100 participants as to how uncanny and how synchronized speech with lip movement was perceived to be. (A full description of the stimuli used in the experiment is provided in the third section.) The results revealed a strong relationship between how uncanny a character was perceived to be and a lack of synchronization between lip movement and speech: those characters with disparities in synchronization were perceived as less familiar and more strange than those characters rated as close to perfect lip-synchronization.

Synchronization problems with the recorded voice for early sound cinema heightened a viewer’s awareness that the figure was not real and was simply a manufactured artifact (Spadoni, 2000, p. 34). A viewer was reminded that figures onscreen were merely fabricated objects created within a production studio. The uncanny was increased as figures were perceived as, “a reassembly of a figure” easily disassembled within a movie theatre (Spadoni, 2000, p. 19). The results from the UM study (Tinwell & Grimshaw, 2010) imply that the implications of asynchrony for speech and the uncanny for human figures within the classic horror cycle of Hollywood film also apply to virtual characters intended for computer games. The zombie characters the Witch and the Tank from the computer game *Left 4 Dead* (2008), received less than average scores for perceived lip-synchronization. The jerky, haphazard movement of the Witch’s lips appeared disparate from the high-pitched cries and shrieks spewed out by this character. As the Witch proceeded to attack, her presence seemed evermore overwhelming as sounds appeared to emulate from an incorporeal and uncontrollable being in a similar manner to Dracula’s laughter noted earlier. Similarly, participants seemed somewhat confused by the chaotic movement and irregular sounds generated by the Tank character making the viewer feel panicked and uncomfortable.

The stimuli for this study were presented in different settings and as different actions. Some were presented as talking heads, for example the
Chatbot character, whilst others moved around the screen, for example the Tank and the Witch. A further study is required to determine the actual causality of lip-synchronization as a significant contributor towards the uncanny when not associated with other factors of facial animation and sound. Thus, we intend a further experiment to test the hypothesis: Uncanniness increases with increasing perceptions of lack of synchronization between the character’s lips and the character’s sound.

At present there are no standards set for acceptable levels of asynchrony for computer games as there are for television. It may well be that these acceptable levels are the same across the two media but it might equally be the case that the interactive nature of computer games and the use of different reproduction technologies and paradigms propose a different standard. For example, perhaps it is the case that current technological limitations in automated lip-syncing tools require a smaller window of acceptable asynchrony for computer games than previously established for television. We hope the future experiment noted above will also ascertain if viewers are more sensitive to an asynchrony of speech for virtual characters where the audio stream precedes video (as has been previously identified for the television broadcasting industry).

**ARTICULATION OF SPEECH**

Hundreds of individual muscles contribute to the generation of complex facial expressions and speech. As one of the most complex muscular regions of the human body, and with increased realism for characters, generating realistic animation for mouth movement and speech is a challenge for designers (Cao et al., 2004; Plantec, 2007). Even though the dynamics of each of these muscles is well understood, their combined effect is very difficult to simulate precisely. Whilst motion capture allows for the recording of high fidelity facial animation and expression, this technique is mostly useful for FMV. Recorded motions are difficult to modify once transferred to a three-dimensional model and the digital representation of the mouth remains an area requiring further modification. Editing motion capture data often involves careful key-framing by a talented animator. A developer may edit individual frames of existing motion capture data for prerecorded trailers and cut scenes yet, for computer games, most visual material is generated in real-time during gameplay. For in-game play, automatic simulation of the muscles within and surrounding the mouth is necessary to match mouth movement with speech. Motion capture by itself cannot be used for automated facial animation.

To create automatic visual simulation of mouth movement with speech, computer game engines require a set of visemes as the visual representation for each phoneme sound. Faceposer (Valve, 2008) uses the phoneme classes phonemes, phonemes strong, and phonemes weak with a corresponding viseme to represent each syllable within the International Phonetic Alphabet (IPA). Prerecorded speech is imported into a phoneme extractor tool that extracts the most appropriate phoneme (and corresponding viseme) for recognized syllables. Editing tools allow for the creation of new phoneme classes, or to modify the mouth shape for an existing viseme.

The UM study (Tinwell & Grimshaw, 2010) identified a strong relationship between how uncanny a character was perceived to be with a perceived exaggeration of facial expression for the mouth. The results implied that those characters perceived to have an over-exaggeration of mouth movement were regarded as more strange. Thus, uncanniness increases with increasing exaggeration of articulation of the mouth during speech. Finer adjustments to mouth shapes using tools such as Faceposer may prevent a perceived over-exaggeration of articulation of speech, yet such adjustments are time consuming for the developer. If no original visual footage is available for speech,
judgements made to correct mouth shapes that appear too strong or too weak are likely to be based on the subjective opinion of an individual developer. Even then, the developer is still constrained by the number of mouth and facial muscles available to modify within the 3D model, which may not include an exhaustive depiction of every single muscle used in human speech.

To avoid the uncanny, working with the range of mouth shapes and facial expression that current technology allows for within tools such as Faceposer, the developer should at least avoid an articulation of speech that may appear over-exaggerated. The mouth shape for the phoneme used to pronounce the word “no” (“n” in Faceposer) may be applicable if the word is pronounced in a strong, authoritative way, but would appear overdone and out of context if the same word was used to provide reassurance in a calming and less domineering manner.

Indeed, if the developer wishes to create an uncanny sensation for a zombie character, adjusting mouth shapes so that articulation of speech appears over-exaggerated may enhance the fear factor for such characters by increasing perceived strangeness. In the same way that a snarling dog or ferocious beast may raise the corners of their mouths to show their teeth in an aggressive way, viewers may be made to feel uncomfortable by overstated mouth movements that suggest a possible threat.

**SUMMARY AND CONCLUSION**

In summary, attributes of speech that may exaggerate the uncanny for realistic, human-like characters in computer games are:

1. A level of human-likeness for a character’s speech that does not match the fidelity of human-likeness for a character’s appearance
2. An asynchrony of speech with lip movement
3. Speech that is of an incorrect pitch or tone.
4. Speech delivery that is perceived as slow, monotone, or of the wrong tempo
5. An over-exaggeration of articulation of the mouth during speech.

Whilst such characteristics of speech may adorn the spine tingling sensation associated with the uncanny for antipathetic characters in the horror genre of games, a developer may risk the uncanny if such characteristics exist for empathetic characters. The protagonist Mary Smith, as featured in the tech demo for the adventure game *Heavy Rain* (2006), may have been intended to evoke affinity and sympathy from the audience. Instead, Mary Smith was regarded as strange and abnormal: Uncanny speech for this character contributed to just such a negative response from the audience. The speech was not only judged as lacking synchronization with lip movement but an inaccurate pitch and lack of human-likeness raised doubt as to whether the voice actually belonged to the character or not. Attributes such as these reduced the overall believability for Mary Smith. However, for zombies such as the Tank and the Witch from the survival horror game *Left 4 Dead* (Valve, 2008), uncanny speech increased (in a desired manner) how strange and freakish these characters were perceived to be.

The outcomes from this investigation show that the majority of characteristics for uncanny speech in computer games may be induced by current technological limitations in the production, reproduction, and control of virtual characters. Restrictions as to the range of facial muscles available to manipulate in automated facial animation tools used to generate footage in real-time is a current constraint for achieving realism in computer games comparable to film. It seems there is a lack of an exhaustive range of mouth shapes to fully represent each phoneme sound and variation of interpolation between syllables in a range of different contexts. Such constraints may contribute to a perceived asynchrony of speech and mouth
shapes being used for syllables that do not accurately convey the prosody or context of speech.

Computer games may always be playing catch-up with the levels of anatomical fidelity achieved in film for facial animation, however developments in procedural game audio and animation may provide a solution for uncanny speech.

As Hug states, the future of sound in computer games is moving towards procedural sound techniques that allow for the generation of bespoke sounds, to create a more realistic interpretation of life within the 3D environment. For in-game play dynamic sound generation techniques, “such as physical modelling, modal synthesis, granulation and others, and meta forms like Interactive XMF” will create sounds in real-time responding to both user input and the timing, position, and condition of objects within gameplay (Hug, 2011).

Using procedural audio (speech synthesis in this case), a given line of speech may be generated over a differing range of tempos using a delivery style appropriate for the given circumstance. For example, the sentence “I don’t think so” may be said in a slow, controlled manner, if carefully contemplating the answer to a question. In contrast, a fast-paced tone may be used if intended as a satirical plosive when at risk of being struck by an antagonist.

Procedural animation techniques for the mouth area may also allow for a more accurate depiction of articulation of mouth movement during speech. Building on the existing body of research into real-time, data-driven, procedural generation techniques for motion and sound (for example, Cao et al., 2004; Farnell, 2011; Mullan, 2011), a tool might be developed that combines techniques for the procedural generation of emotive speech in response to player input (actions or psychophysiology) (Nacke & Grimshaw, 2011) or game state. Interactive conversational agents in computer games or within a wider context of user interfaces may appear less uncanny if the tempo, pitch, and delivery style for their speech varies in response to the input from the person interacting with the interface. Such a tool will aid in fine-tuning the qualities of speech that will, depending on the desired situation, reduce or enhance uncanny speech.

REFERENCES


KEY TERMS AND DEFINITIONS

**Audio-Visual:** An artifact with the components image and sound.

**Cross-Modal:** Interaction between sensory and perceptual modes, in this case, of vision and hearing.

**Realism:** Representation of objects as they may appear in the real world.

**Uncanny Valley:** A theory that as human-likeness increases, an object will be regarded as less familiar and more strange, evoking a negative effect for the viewer (Mori, 1970).
Virtual Character: A digital representation of a figure onscreen.

Viseme: A visual representation of a mouth shape for a particular speech utterance such as “k,” “ch” and “sh.” Those with hearing impediments can use visemes to lip read and understand the spoken language when unable to hear sound.

ENDNOTE

1 In the field of psychoacoustics, synchrony and synchresis are closely related to the ventriloquism effect.