Preliminary version of: **Qu, C., Brinkman, W. P., Ling, Y., Wiggers, P., & Heynderickx, I. (2013). Human perception of a conversational virtual human: an empirical study on the effect of emotion and culture. *Virtual Reality*, 1-15. DOI:** 10.1007/s10055-013-0231-z

# Human perception of a conversational virtual human: An empirical study on the effect of emotion and culture

Chao Qu[1]

Delft University of Technology, Mekelweg 4, 2628 CD Delft, the Netherlands

Willem-Paul Brinkman

Delft University of Technology, Mekelweg 4, 2628 CD Delft, the Netherlands

Yun Ling

Delft University of Technology, Mekelweg 4, 2628 CD Delft, the Netherlands

Pascal Wiggers

Delft University of Technology, Mekelweg 4, 2628 CD Delft, the Netherlands

Ingrid Heynderickx

Delft University of Technology, Mekelweg 4, 2628 CD Delft, the Netherlands

Philips Research Laboratories, High Tech Campus 34, 5656 AE Eindhoven, the Netherlands

[1] Correspondence to aquchaos@gmail.com
Tel.:+31681808384

Preliminary version of: **Qu, C., Brinkman, W. P., Ling, Y., Wiggers, P., & Heynderickx, I. (2013). Human perception of a conversational virtual human: an empirical study on the effect of emotion and culture. *Virtual Reality*, 1-15. DOI:** 10.1007/s10055-013-0231-z

# Abstract

Virtual reality applications with virtual humans, such as virtual reality exposure therapy, health coaches, and negotiation simulators, are developed for different contexts and usually for users from different countries. The emphasis on a virtual human's emotional expression depends on the application; some virtual reality applications need an emotional expression of the virtual human during the speaking phase, some during the listening phase and some during both speaking and listening phases. Although studies have investigated how humans perceive a virtual human's emotion during each phase separately, few studies carried out a parallel comparison between the two phases. This study aims to fill this gap, and on top of that, includes an investigation of the cultural interpretation of the virtual human's emotion, especially with respect to the emotion's valence. The experiment was conducted with both Chinese and non-Chinese participants. These participants were asked to rate the valence of seven different emotional expressions (ranging from negative to neutral to positive during speaking and listening) of a Chinese virtual lady. The results showed that there was a high correlation in valence rating between both groups of participants, which indicated that the valence of the emotional expressions was as easily recognized by people from a different cultural background as the virtual human. In addition, participants tended to perceive the virtual human's expressed valence as more intense in the speaking phase than in the listening phase. The additional vocal emotional expression in the speaking phase is put forward as a likely cause for this phenomenon.

**Keywords:** virtual reality, virtual human, emotion, affective computing, culture

# 1. Introduction

To create a feeling of being "present" in virtual reality is essential to the success of many virtual reality applications such as training (Broekens et al., 2011), coaching (Rizzo et al., 2011), therapy (Brinkman et al., 2012), and games (Isbister, 2006). A feeling of being "present" in virtual reality may be achieved by making the virtual reality environment as natural as possible. Human-computer interaction, including human-virtual human interaction, is inherently natural and social (Reeves & Nass, 1996), and so, is an essential component in the realism of the virtual environment. Without proper behavior of the virtual human, users may not be able to "suspend disbelief" and the effectiveness of the virtual reality application will decrease.

Considering the importance of emotion in human-human communication, emotion may also help people to establish a better relationship with virtual human (Reeves & Nass, 1996). As Picard, Vyzas, and Healey (2001) argues, without some emotional skills, machines will not appear intelligent when interacting with people. Therefore, multiple technologies to give virtual human the abilities of generating human acceptable expressions have been developed in recent decades (Ersotelos & Dong, 2008).

Different applications require different levels of emphasis on how the virtual human express their emotions. Even when implemented in only part of the application, emotional expressions can be effective. In a health coach application, for example, the virtual human might mainly need to speak to motivate the user, and emotional expressions during speaking are most important. In a virtual reality exposure therapy for fear of public speaking, the virtual human only needs to listen, and so, emotional expressions while listening are most important. In some applications, such as for a role playing game or a negotiation simulator, the full range of speaking and listening is used and might benefit from emotional expressions. Studies on generating and evaluating the emotional agents normally only focus on either listening (Slater, Pertaub, & Steed, 1999; Wong & McGee, 2012) or speaking (MacDorman, Coram, Ho, & Patel, 2010; Qiu & Benbasat, 2005). Studies that do include both speaking and listening,(e.g., Core, Traum, Lane, and Swartout (2006), Broekens, Harbers, et al. (2012), Link, Armsby, Hubal, and Guinn (2006)) focus mainly on the conversation and communication as a whole, and do not separately investigate the speaking and listening phase of the whole conversation. To our knowledge, no study has directly compared the impact of emotional expressions during speaking and listening in virtual reality. In the current study, the virtual human's valence state was manipulated while she was speaking and listening from negative to positive, and the impact on the participants' perception was examined.

Besides the difference between listening and speaking, culture might also be an important factor for a designer to consider as many applications are used all across the world nowadays. Especially for some virtual reality applications, such as virtual reality exposure therapy for patients with social phobia (Brinkman et al., 2012), it is crucial to understand how people with a different cultural background perceive the affective behavior of virtual humans. Several studies have already focused on the effect of cultural differences on evaluating virtual human's emotions. For example, Jack, Garrod, Yu, Caldara, and Schyns (2012) showed that facial expressions of emotion are culture specific. However, Yun, Deng, and Hiscock (2009) found that cultural background has little effect on emotion perception. Kleinsmith, De Silva, and Bianchi-Berthouze (2006) evaluated cultural impact on perception of emotion and found that emotions are both universal and culturally specific. Therefore, similar to the studies for perceiving emotional expressions of real humans, the universality of emotion perception of virtual human seems also still inconclusive. In addition, most research is only limited to the investigation of head-only virtual human with facial expressions and far less research is devoted to emotional expressions from a 3D virtual human which expresses its emotional state also via gaze, head movement or voice intonation.

In summary, this study involves two research questions: (1) whether emotional expressions of virtual human are perceived differently depending on the cultural background of the perceiver, and (2) whether a person is more perceptive to emotional expressions in one of the two phases (the speaking phase and listening phase) or whether a person treats these two phases as equally important when rating the virtual human's emotions? To answer these research questions, we designed a virtual human representing a Chinese lady at an age around 25. She had the ability to show multiple emotional states in multiple non-verbal and verbal ways: i.e., through facial expression, head movement, gaze and voice intonation. During the listening phase, the virtual human's emotional behavior was expressed by non-verbal communication only, while during the speaking phase, the emotional behavior was expressed by both verbal (i.e., intonation) and non-verbal communication. To avoid a possible emotional bias from the content of the conversation, a relatively neutral topic, i.e. conference attendance, was selected in this experiment. Petrushin (1999) pointed out that humans are not perfect in decoding manifest emotions such as anger and happiness in voice intonation only. Therefore, as a first step, only three basic emotional valence states (positive, neutral and negative) were used in this study. In order to test the effect of cultural influence on the perception of the emotions expressed by the virtual human, two groups of participants were recruited: from the same culture as the virtual human and from other cultures. We chose to compare Chinese versus non-Chinese participants, because it is known from cultural models (Hofstede, 2001) that the difference in cultural values is significant between these two groups, and since two of the authors experienced these differences while living in Europe. Moreover the background of these authors facilitated the recruitment of Chinese participants.

Based on knowledge available in the literature, we envision the following two hypotheses related to our research questions.

**HYPOTHESIS 1**: Individuals with the same cultural background as the virtual human perceive the valence state of the virtual human differently from individuals with a different cultural background.

Especially, as the virtual human was speaking Chinese, participants with a different cultural background could not understand what the virtual human said during verbal communication. Hence, participants with a different cultural background from the virtual human are expected to perceive her emotion differently from participants with the same cultural background.

**HYPOTHESIS 2**: The virtual human's expressed valence is perceived as more intense in the speaking phase than in the listening phase.

Since the speaking phase also allows including verbal expression of the emotion, it seems likely that compared to the listening phase, the emotion expressed in the speaking phase is perceived as more intense.

The rest of the paper is structured as follows. Section 2 provides theoretical background on how a virtual human can express emotion through facial expression, gaze, head movements, and voice intonation. In addition, it discusses cultural differences in emotion recognition and various emotional models, needed to understand the rest of the paper. Section 3 provides a description of the apparatus, validation of the stimuli material and the procedure of the experiment, and its results are presented in section 4. Finally, in section 5 the findings of the study are discussed and conclusions are drawn.

# 2. Theoretical Background

No matter what roles virtual humans play in a virtual world, they need to elicit an anthropomorphic interaction with their human users. This requires vast knowledge of various human aspects including facial expression, gaze, head movement, voice expression and their cultural difference in order to make the virtual human believable, responsive and interpretable.

## 2.1 Facial Expression of a Virtual Human

Facial expression is one of the options to express human emotion, and as such plays a substantial role in depicting human characters. Started in the early 70s 80s (Parke, 1972; Platt & Badler, 1981), face modeling and animation have been a continuous research topic for many years. From early 2000s, more flexible emotion representations were created with MPEG-4-based facial animation (Tsapatsoulis, Raouzaiou, Kollias, Cowie, & Douglas-Cowie, 2002). Recent advances in facial animation that allow to produce a rich set of effects on synthetic humans already had their impact on the industry (Ersotelos & Dong, 2008).

Multiple approaches have been proposed to create naturally looking facial expressions; they can be categorized as follows: (1) simulation or physically based models, which try to model the anatomical structure of the face as well as the underlying dynamics (Kahler, Haber, & Seidel, 2001; Y. Lee, Terzopoulos, & Walters, 1995; Waters, 1987), (2) performance driven models, which reassemble frames from video footage or motion capture data of a real person to yield the desired facial expression (Brand, 1999; Bregler, 1997; Chuang & Bregler, 2002; Ezzat, Geiger, & Poggio, 2004; Litwinowicz & Williams, 1994), and (3) parameterized based models, which assign weights to the vertices of meshes representing the face, such that during animation the vertices are moved according to the weights (Cohen & Massaro, 1993; Parke, 1974; Zhang, Liu, Quo, Terzopoulos, & Shum, 2006). Considering the high computational load required for the simulation or physically based models and the high costs for the motion capture equipment needed for performance driven models, we decided to choose an easily repeated facial expression animation based on a parameterized model for this study.

## 2.2 Head Movement and Gaze of a Virtual human

Besides facial expressions, also head and eye movements were implemented in the virtual human used in our experiment. Head movements and eye gaze are two important sources of emotional feedback in interaction (Cassell & Thorisson, 1999; J. Lee & Marsella, 2012; Ruttkay & Pelachaud, 2005). They are essential to embody interactive conversational systems (Cassell et al., 1994) and it is relatively simple to create primarily nods and glances towards or away from the user. Still the correct timing is essential (Cassell & Thorisson, 1999). Research of Lance, Rey, and Marsella (2008) and J. Lee, Prendinger, Neviarouskaya, and Marsella (2009) show how head movements and gaze can be embedded into a virtual character.

# 2.3 Voice Expression of a Virtual human

Along with the non-verbal emotional expressions, emotion can also be expressed by voice intonation when the virtual human is talking. Speech was once considered as the main channel to carry most, or even all, the necessary information in a conversation (Ochsman & Chapanis, 1974). This idea has been countered by a growing body of research on believable, life-like embodied conversational agents (Bates, 1994). Still, the importance of the voice in emotion expression cannot be denied (Scherer, 1995). Many studies have investigated emotional effects in voice and speech (Bailenson, Yee, Merget, & Schroeder, 2006; Petrushin, 1999; Scherer, 2003), and emotion expressed in the voice of virtual humans (Cerezo & Baldassarri, 2008; Moridis & Economides, 2012). The intonation of the voice was therefore also considered as an important aspect of the virtual human’s emotional expression in this study.

# 2.4 Cultural Difference

Culture, like age, gender, posture and context, is one of the many factors affecting emotion expression (Picard, 1998). A long-time question in the study of human emotion is the extent to which emotional expressions are universal or culturally determined (Elfenbein, Beaupre, Levesque, & Hess, 2007). Cultural background may influence the rate of emotion recognition (Matsumoto, 2002). When an expresser of an emotion and the perceiver of the emotion have the same cultural background, the perceiver's recognition rate is found to be higher than when the expresser and perceiver have a different cultural background (Elfenbein, 2003; Elfenbein & Ambady, 2002; Elfenbein et al., 2007). However, Darwin (1872) and Tomkins (1962) (1963) argue that universal emotions do exist, studies also show universality in the facial expression of emotion and its perception, and attribute only little effect of cultural background on emotion perception from facial expressions (Ekman, 1994; Ekman & Friesen, 1971; Ekman et al., 1987; Matsumoto, 2002, 2007).

The question of impact of cultural background can be extended to human-virtual human interaction. Although various studies show that people can correctly identify emotions expressed by embodied agents in general (Bartneck, 2001; Schiano, Ehrlich, Rahardja, & Sheridan, 2000), how good this performance is retained in different cultures needs to be considered. Clear indications support the statement that culture can shape the expression and interpretation of emotions (Keltner & Ekman, 2000). Culture as a factor has also been studied in the interaction with computers. For example, Dotsch and Wigboldus (2008) and Brinkman et al. (2011) have found a difference in emotional reaction to a virtual human with ethnic appearance that match or did not match the person's ethnicity. Endrass, Rehm, and Lipi (2011) show that in German and Japanese cultures, the user's perception of an agent conversation can be enhanced by a culturally prototypical performance of gestures and body postures. Kleinsmith et al. (2006) worked on the cross-cultural difference of recognizing affect from virtual human's body posture and suggest to consider culture as one specific factor for the implementation of agents. Meanwhile Jan, Herrera, and Martinovski (2007) mention that in Arabian and US American cultures, gaze, proximity and turn-taking behavior are all culture related. These results reveal that participants perceive behavior that is in line with their own cultural background differently from behavior that is typical for a different cultural background. In the work presented in this paper, cultural background is considered as a variable which is expected to influence how people perceive the emotional expression of the virtual human.

## 2.5 Dimensional Emotion Model

Although for facial expressions six universal basic emotions exist (Ekman, Rolls, Perrett, & Ellis, 1992), for language people's categorization of verbal labels to describe their everyday life emotions vary between languages and cultures (Russell, 1991). Instead of placing these expressed emotions in categories, i.e. a discrete emotional approach, others suggest placing them in a multi-dimensional space, i.e. a dimensional approach (Fox, 2008). Three broad dimensions have often been proposed to describe affect (Mehrabian & Russell, 1974): i.e. valence, arousal and dominance. Valence is variously referred to as positive and negative affect or as pleasant and unpleasant feelings. The arousal dimension ranges emotions from deep sleep to frenetic excitement. Dominance focuses on the expression of social control and aggression, and varies between submissive and dominant (Schroder, 2004). Compared to the discrete emotional approach, the dimensional approach often uses subjective reports of feelings as its main dependent variable. As such, it has a strong empirical base. Support for the existence of these dimensions has come from research into subjective reports, physiological responses, neural circuits, and cognitive appraisal (Barrett, 2006; Fox, 2008). Furthermore, Wierzbicka (1995) and Church and Katigbak (1998) also investigated the cross-cultural universality of the emotional dimensions. Their results showed the universality of the valence and arousal dimensions. The study presented in this paper focuses on the valence dimension only. Although participants were asked to rate the virtual human's emotion on all the three dimensions, only the valence dimension was used for data analysis.

# 3. Experiment

## 3.1 Participants

Twelve Chinese (7 female and 5 male) and twelve non-Chinese (5 female and 7 male) students from the Delft University of Technology participated in the experiment. Their age ranged from 24 to 38 years with a mean of 27.8 (*SD* = 3.4) years. All participants were naive with respect to the hypotheses. Written informed consent forms were obtained from all the participants. The experiment was approved by the university ethic committee.

## 3.2 Creating the virtual human

Although Cowell and Stanney (2003) found that users generally prefer to interact with a youthful character matching their ethnicity, they found no significant preference for character gender. Furthermore Kulms, Kramer, Gratch, and Kang (2011) showed that actual behavior and its evaluation are more important for the evaluation than gender stereotypes. Therefore, a Chinese virtual lady aged around 25 years was specially created for this study.

The model of the virtual human was created by FaceGen and 3Ds MAX. All main factors which were considered to contribute to emotion expression were combined; the virtual human's facial expression, her head and eye movements and her voice intonation were manipulated to express emotion during the conversation. To create facial expressions, an easily repeated facial expression animation method was used. This method rigged the face mesh into 22 action units with 18 features (Gratch et al., 2002), where each feature was an anchor point attached to a set of vertices of the face. A model for the face dynamics that was able to control the intensity of the expression, its onset, peak and decay was defined. As such, the virtual human had the ability to show any intensity and any combination of the six basic Ekman facial expressions (Ekman, Friesen, & Hager, 2002). The validation of this approach was shown by Broekens, Qu, and Brinkman (2012). By setting the values for the three emotional dimensions (i.e., the valence, arousal and dominance), and for the expression duration, any emotion could be expressed by the virtual human. The facial expressions from neutral to negative or from neutral to positive, used by the virtual human in our experiment are shown in Figure 1.

The participants were asked to judge the emotional state of the virtual human, and so, there was no interaction between the participant and the virtual human. The participant was told that the scene contained a virtual lady talking with a human, but that the human voice was removed. Therefore, problems related to timing (i.e., whether the virtual human should or should not show an expression at a certain point of time) were avoided, and the participant could focus on the emotional behavior of the virtual human herself.

Seven conditions were included in the experiment, all varying in the emotional states of the virtual human. Since the scenario was conversation based, two continuously alternating phases could be identified, i.e. one in which the virtual human was speaking and one in which she was listening. These phases allowed the virtual human to express her emotion differently in the two phases. In the speaking phase, the virtual human used voice and non-verbal communication to express her emotions, while in the listening phase the virtual human only used non-verbal communication to express her emotions. Three emotional states were created

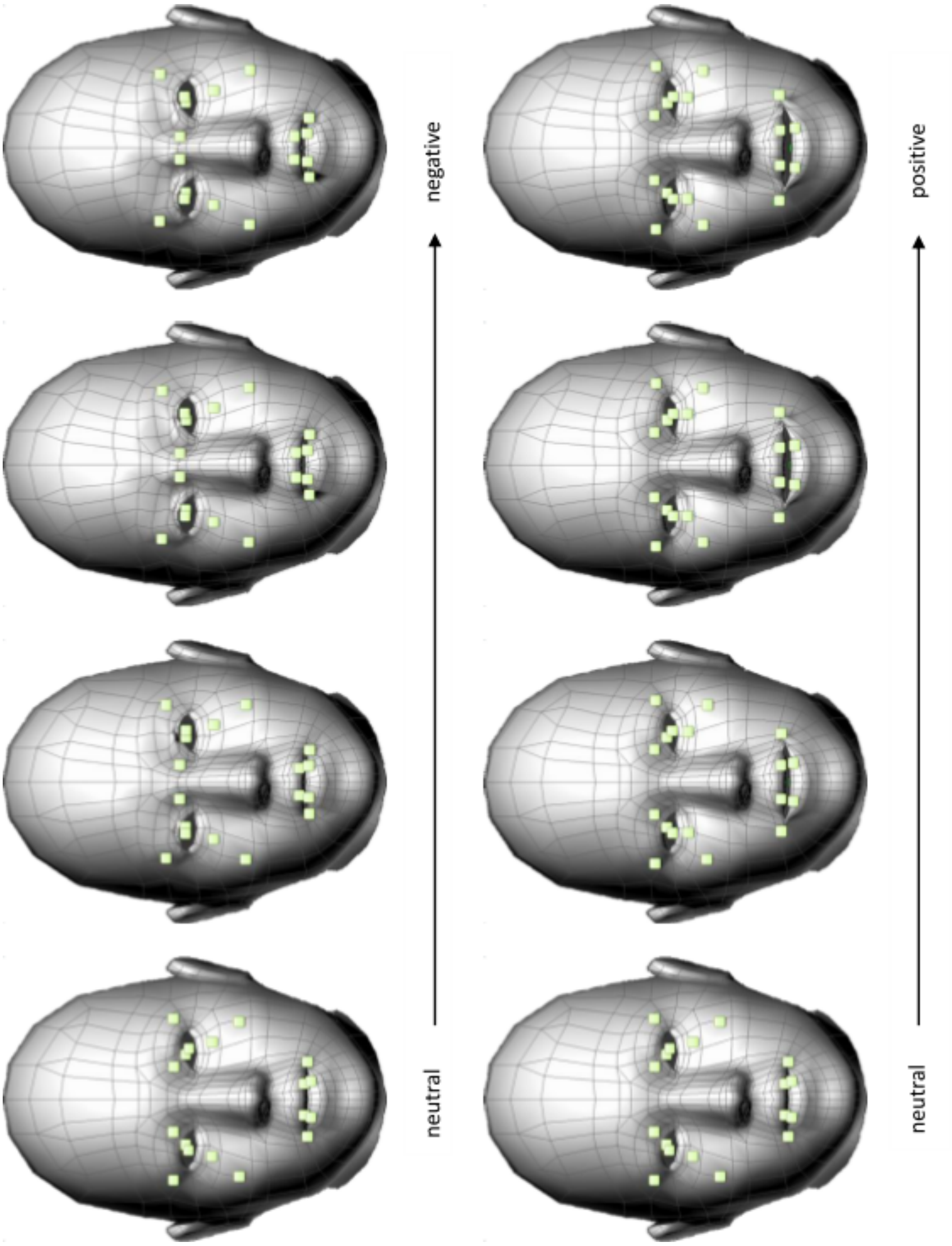


Figure 1: Emotions expressed by moving some action units (i.e., the small squares in the figure) of the face mesh. Left column: emotions changing from neutral to negative; right column: emotions changing from neutral to positive.

for both phases (i.e., positive, neutral and negative state), and they formed the basis for the seven different conditions, shown in Figure 2. As the combination of positive (negative) listening and negative (positive) speaking included contradictory emotional information of the virtual human in the speaking and listening phase, these combinations were considered unnatural, and so, were excluded from the experiment. Taking the neutral attitude in both the speaking and listening phase as the baseline, it was expected that participants would give a higher valence score when the virtual human responded positively either in the listening or speaking phase. Assuming that there would be no interaction between the speaking and listening phase and that both phases would have a similar impact on the expressed valence intensity, the seven conditions could be ordered into five groups: highly negative (*S-L-*), lowly negative (*S-L0*, *S0L-*), neutral (*S0L0*), lowly positive (*S+L0*, *S0L+*) and highly positive (*S+L+*). If the intensity of the expressions with a negative or positive valence would be equal, these five groups could be projected on a single valence scale as is done in Figure 2 (shown as the predicted valence value axis). Comparing the actual valence values obtained in the experiment to the predicted valence values would make it possible to study hypothesis 2 about the experience of the valence intensity in the two phases of the conversation.

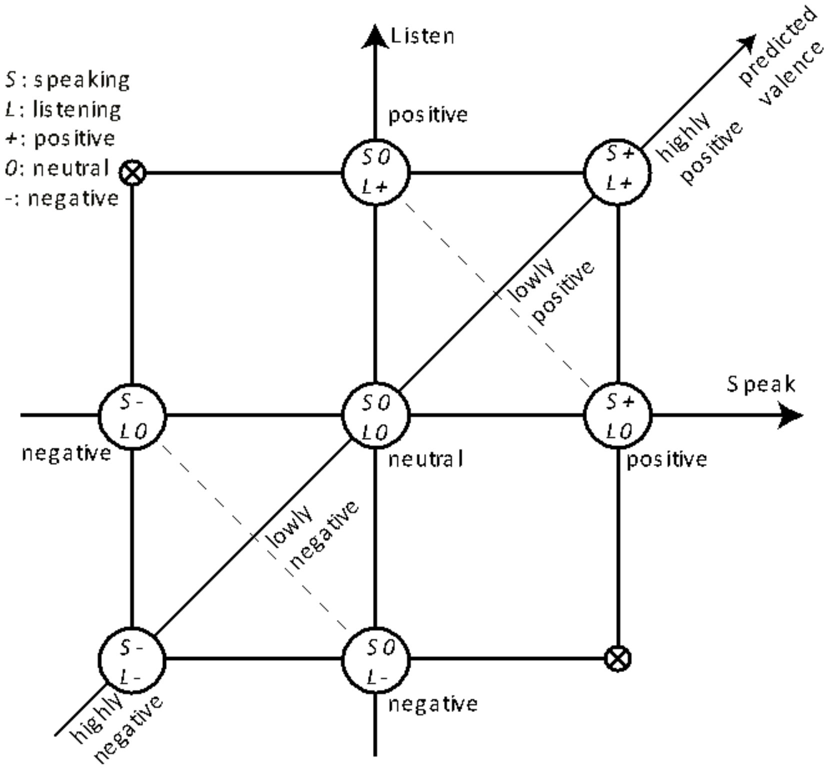
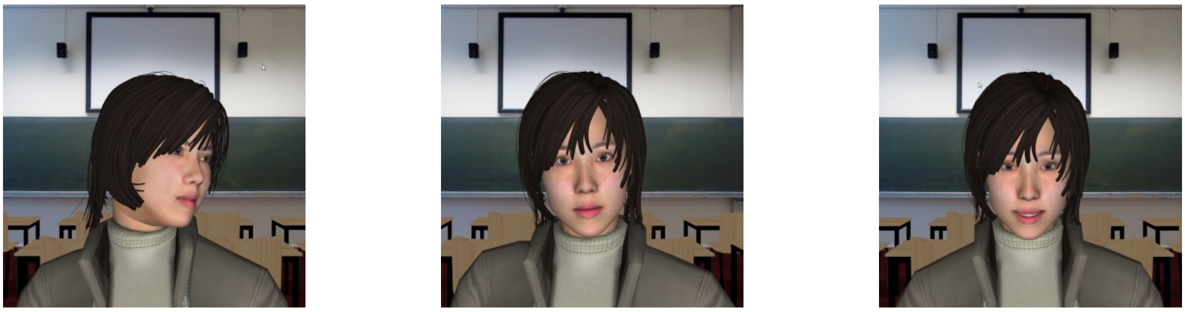


Figure 2: Seven conditions, existing of combinations of an emotional state in the speaking and listening phase of a conversation, as used in the experiment and their corresponding predicted valence intensity

The participants were asked to sit in front of the virtual human (displayed only above her chest on a computer screen), right at the place where the virtual human's "conversational partner" would sit. With this set up, the participants could well perceive the virtual human's emotional state, expressed by her vocal expression, facial expression, eyes and head movements. When expressing a positive emotional state, the virtual human would show a happy facial expression, and once in a while would nod her head to agree with her conversation partner. Her eyes would mainly look at her conversational partner, only occasionally look away (Figure 3c). When expressing a negative emotional state the virtual human would have an angry facial expression and would continuously look away showing limited interest in her conversation partner (Figure 3a). The intensity of both the positive (happy) and negative (angry) emotional expression was evaluated in a previous study (Broekens, Qu, et al., 2012) to ensure that they both could be identified by individuals. The neutral expression was the default facial expression of FaceGen, with the six Ekman basic emotion (Ekman et al., 1992) parameters set to zero and with all other morph modifiers removed when generating the face model.



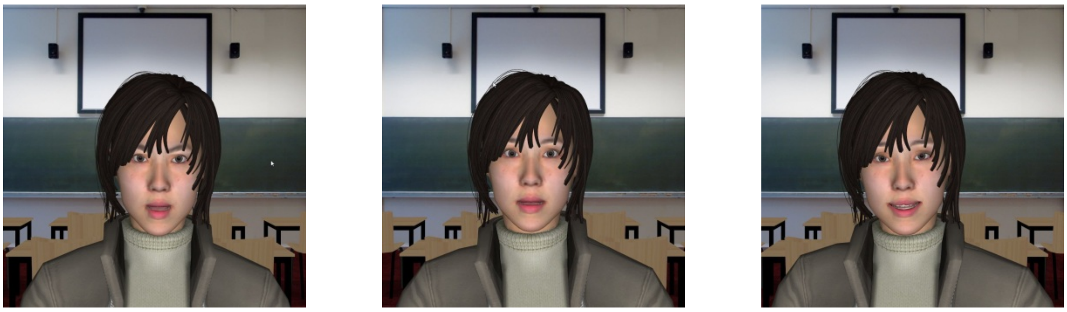
(a) Negative: angry facial expression, only looking at her conversation partner at the beginning, gradually losing interest and starting to look around.

(b) Neutral: neutral facial expression while constantly looking at her conversation partner with slight eye movements.

(c) Positive: happy facial expression while constantly looking at her conversation partner, showing some slight eye movements, and occasionally nodding her head.

Figure 3: Different emotional states of the virtual human in her listening phase

In the speaking phase, the virtual human would look directly at her conversation partner. In the negative speaking condition she would have a negative facial expression (Figure 4a), while in the positive speaking condition she would have a positive facial expression (Figure 4c). In addition, speech with either a negative or positive intonation was added to the virtual human.



(a) Negative: angry facial expression while looking at her conversation partner, and speaking with a negative voice intonation.

(b) Neutral: neutral facial expression while constantly looking at her conversation partner, and speaking with a neutral voice intonation.

(c) Positive: happy facial expression while constantly looking at her conversation partner, and speaking with a positive voice intonation

Figure 4: Different emotional states of the virtual human in her speaking phase

# 3.3 Emotion Validation

As mentioned already above, for the speaking phase, verbal communication was added to the virtual human. The voice of the virtual human was recorded in Chinese by a Chinese linguistics student. Her voice was recorded 3 times, each time expressing a different emotional state: positive, neutral and negative. A small separate study, in which 6 Chinese participants, 3 male and 3 female with an average age of 27 (*SD* = 0.5) years, were asked to rate the valence of the recorded voice on a scale from 1 (negative) to 9 (positive), showed that the emotion in the recorded voice was indeed perceived as intended, *F*(2,10) = 25.29, *p* < .001. The negative voice was significantly lower than the neutral voice, *t*(5) = 3.87, *p* = .012, and the positive voice, *t*(5) = 6.52, *p* < .001. Further, the positive voice was significantly higher than the neutral voice, *t*(5) = 3.61, *p* = .015. The means and standard deviations of the scores on the positive, the neutral and the negative voice were *M* = 7.8, *SD* = 1.9; *M* = 5.7, *SD* = 2.0; *M* = 1.7; *SD* = .8, respectively.

Making a fair comparison between the listening and speaking phase requires that the intensity of the non-verbal communication is similar in both phases. For example, the virtual human's facial and body expression in the lowly negative speaking phase and lowly negatively listening phase (see Figure 2) should have a similar impact on the valence intensity. To test this, an additional small study was conducted. In this study twelve participants, 5 male and 7 female with an average age of 27 years (*SD* = 1.8) were presented simultaneously with two video clips of the virtual human including both the listening and speaking phase. Half of the participants were Chinese. The participants were asked to rate how easily they could see the difference between the two videos on a scale from very easy (0) to very difficult (100). The participants were explicitly asked not to rate the valence, but only the easiness with which differences were perceived, representing the intensity of the emotion. The videos were presented without sound. The participants were asked to rate 12 pairs in total (*S-L0*/*S0L0*, *S0L-*/*S0L0*, *S+L0*/*S0L0*, *S0L+*/*S0L0*, *S-L-*/*S+L+*, *S-L-*/*S0L0*, *S+L+*/*S0L0*, *S0L0*/*S0L0*, *S+L0*/*S+L0*, *S-L0*/*S-L0*, *S0L+*/*S0L+*, *S0L-*/*S0L-*), presented to each participant in a different random order. Before they rated the pairs, the participants were shown all the possible behaviors of the virtual human so that they could establish an overall frame of reference.

The first step of the analysis was to see whether the more intense stimuli were easier to distinguish from the neutral reference video (*S0L0*) and whether the positive and negative videos were equally distinctive. Therefore, a MANOVA with repeated measures was conducted with the intensity of the video stimuli (high versus low intensity) and the valence direction (positive versus negative) as independent variables. The analysis was conducted on the rating for highly positive (*S+L+*/*S0L0*) and negative (*S-L-*/*S0L0*) videos, and the mean rating for lowly positive (*S+L0*/*S0L0* and *S0L+*/*S0L0*) and negative (*S-L0*/*S0L0* and *S0L-*/*S0L0*) videos across the speaking and listening phase. The analysis found a significant main effect (*F*(1, 11) = 21.91, *p*. = 0.001) for intensity, in that the highly positive or negative videos (*M* = 32, *SD* = 17) were rated as easier to be distinguished than the lowly positive or negative videos (*M* = 44, *SD* = 15). Also a significant (*F*(1, 11) = 15.63, *p*. = 0.002) main effect was found for direction. The positive videos (*M* = 25, *SD* = 15 ) were rated as more easily to be distinguished from the neutral video than the negative videos (*M* = 50, *SD* = 23). The analysis found no significant (*F*(1, 11) = 1.60, *p*. = 0.23) two-way interaction effect, which suggests that the two main effects were constant across the conditions.

The next analysis focused on the question whether, compared to the neutral reference video, the positive or negative differences in the listing or speaking phase were equally distinguishable, and whether this was the same for the positive and negative videos. Therefore, a second MANOVA with repeated measures was conducted with the valence direction and the phase (speaking versus listening) as independent variables. The analysis used the rating for lowly positive speaking (*S+L0*/*S0L0*) and lowly positive listening (*S0L+*/*S0L0*) phase, and the rating for the lowly negative (*S-L0*/*S0L0*) speaking and lowly listening (*S0L-*/*S0L0*) phase. The analysis again revealed that the positive videos (*M* = 28, *SD* = 16) were significantly (*F*(1, 11) = 16.91, *p*.= 0.002) rated as more easily to be distinguished than the negative videos (*M* = 59, *SD* = 24) from the neutral reference video. No significant difference was found between the listening and speaking phase (*F*(1,

11) = 0.14, *p.* = 0.71), and also no significant two-way interaction effect was found (*F*(1, 11) = 0.44, *p.* = 0.52). Figure 5 shows the videos with their predicted valence and the estimated valence. The latter is the *z*-score of the rating for the video subtracted from the rating of the neutral reference video (*S0L0/S0L0*) whereby the rating of negative videos was multiplied by -1. Both the two lowly negative and the two lowly positive video are positioned closely together. In other words the intensity of the non-verbal communication seems similar in the listening and speaking phase. Furthermore, because of the significant difference in rating between negative and positive videos, the neutral reference video seems to be positioned closer to the negative videos than to the positive videos. As illustrated in Figure 5 the predicted and estimated valence values for the videos do not follow a linear function, but rather a cubic function. By using a fitted inverted cubic function, the intensity weighted predicted valence values for the videos were calculated from the estimated valence values, thereby creating values of intended valence intensity to be compared with the perceived valence rating of videos later in the paper.

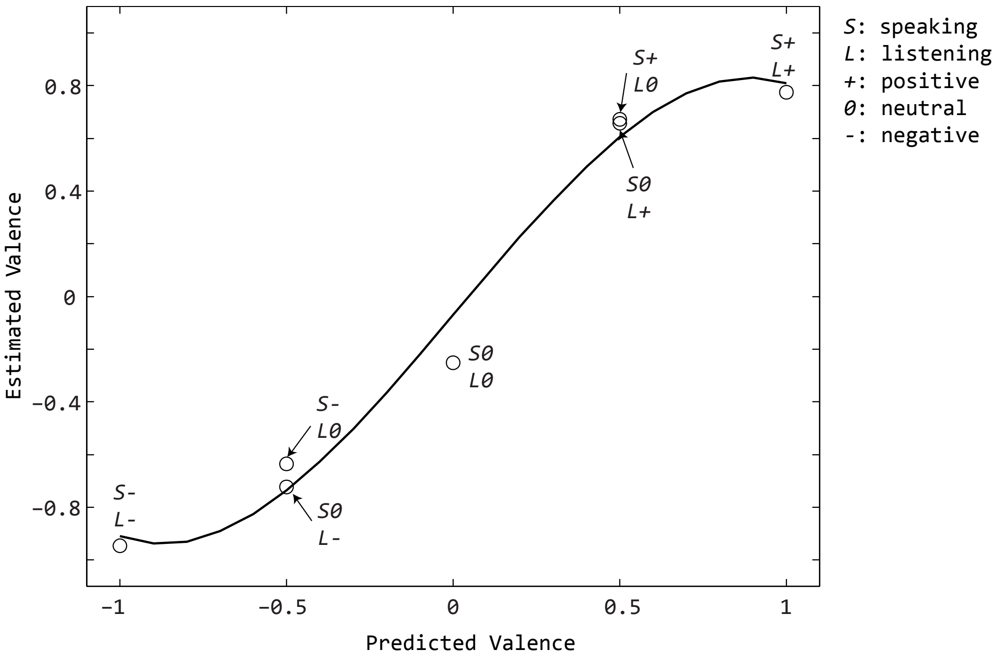


Figure 5: Predicted valence plotted against the estimated valence fitted with a cubic function.

# 3.4 Measurements

There are various ways to quantitatively measure the three emotional dimensions (i.e., valence, arousal and dominance). To ensure the reliability of the emotion measurement, two subjective self-reporting instruments were included in this study: the Self-Assessment Manikin Questionnaire (SAM) (Lang, 1995) and the AffectButton (AFB) (Broekens & Brinkman, 2009).

The SAM questionnaire consists of a series of manikin figures to judge the affective quality (Figure 6). As a nonverbal rating system, the SAM questionnaire represents the intensity value of the three dimensions of emotion: valence, arousal and dominance (Lang, 1995). The first row of SAM manikin figures ranges from unhappy to happy on the valence dimension. The second row represents the arousal dimension, ranging from relaxed to excited. The last row ranges from dominated to controlling, representing the dominance dimension. When instructed on how to use the SAM questionnaire according to the detailed explanation, provided in the instruction manual of Lang, Bradley, and Cuthbert (2008), participants can select one of the nine figures on each row to express their feelings about the emotional stimulus. The manikin figures were taken from the PXLab (Irtel, 2007). Various studies show that the SAM questionnaire accurately measures emotional reactions to imagery (Lang, Bradley, & Cuthbert, 1999; Morris, 1995), sounds (Bradley & Lang, 2007), robot gesture expression (Haring, Bee, & Andre, 2011), etc.

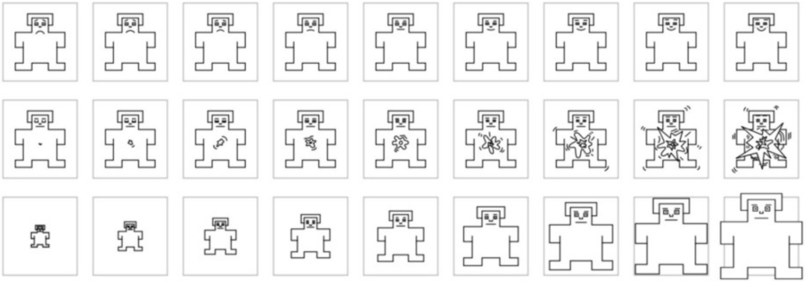


Figure 6: Self-Assessment Manikin Questionnaire, three rows representing the valence, arousal, and dominance dimension respectively[2]

The AffectButton (AFB) offers a flexible and dynamic way to collect users' explicit affective feedback (Broekens & Brinkman, 2009, 2013). The AFB is a button like input interface (Figure 7). In essence, the AFB can be regarded as a navigation tool through a large set of facial expressions. The user can freely move the cursor over the face to change its affective state. Similar to the SAM questionnaire, the AFB returns feedback on the valence, arousal and dominance dimensions. Designed with the intention to be a quick and user-friendly explicit emotion measurement instrument, the reliability and validation of the AFB have been

[2] Copyright © 2001-2006, Hans Irtel. Distributed under the MIT License as certified by the Open Source Initiative.

studied on measuring emotional reactions to words, feelings and music (Broekens & Brinkman, 2009; Broekens, Pronker, & Neuteboom, 2010).

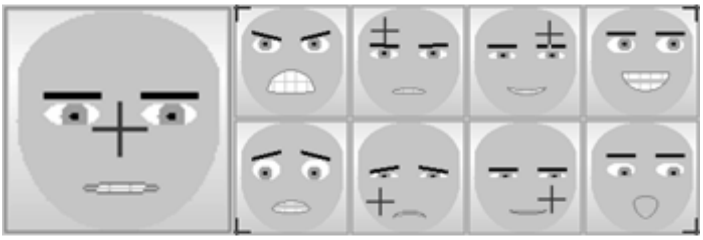


Figure 7: AffectButton and its different appearances while moving the cursor (the cross)

## 3.5 Procedure

Prior to the experiment, participants were provided with an information sheet, and the procedure was explained to them. They were then asked to sign an informed consent form. The experiment was setup as a within-subject design, comprising seven conditions with different emotional expressions both in the listening and speaking phase. In each condition, the participants were asked to watch a short clip (around 1 minute) of a conversation about going to conferences between a Chinese virtual lady and a person. In each clip, the virtual human spoke 10 sentences in total, and was silent in between each sentence, listening to her conversational partner talking. The total length of the virtual human's speaking phase was around 15 seconds, and the rest of the 45 seconds was counted as the virtual human's listening phase. The conversation was in Chinese and the participants could hear what the virtual human said during the speaking phase; during the listening phase, there was no sound of the virtual human's conversational partner. The participants were asked to rate the virtual human's emotional state using both SAM and AFB when they finished watching a clip. The order in which the video clips were shown was randomized across the participants.

# 4. Results

The experiment had seven conditions (Figure 2), with two different measurements and two groups of participants (Chinese and non-Chinese). The data recorded by the SAM questionnaire were integers ranging from 0 to 8, while the data recorded by the AFB were floating-point numbers ranging from -1 to 1. To compare these two measurements, the data were first normalized into z-scores per measurement for each participant across the seven conditions.

The means for the SAM questionnaire and AFB on the valence emotional dimension are shown in Figure 8. A repeated-measures MANOVA was conducted to test the difference between SAM and AFB scores thereby using condition, type of measurement and cultural background as three independent variables, and the z-scores on valence as dependent variable. The analysis also included all two-way and three-way interactions. The results showed no significant difference between SAM and AFB measurement, $F(1,22) = 1.30$, $p = .26$, and also no significant interaction effect.

To test the relationship between these two measurements, a correlation analysis between SAM and AFB scores on the valence dimension was performed. The average scores across all participants for the seven conditions were used. The results showed that SAM and AFB were highly correlated on the valence dimension ($r(7) = 0.995$, $p < .001$). The valence scores collected by these two measurements could therefore be regarded as consistent. This made it possible to only focus on the average of the SAM and AFB z-scores in the remaining analyses.

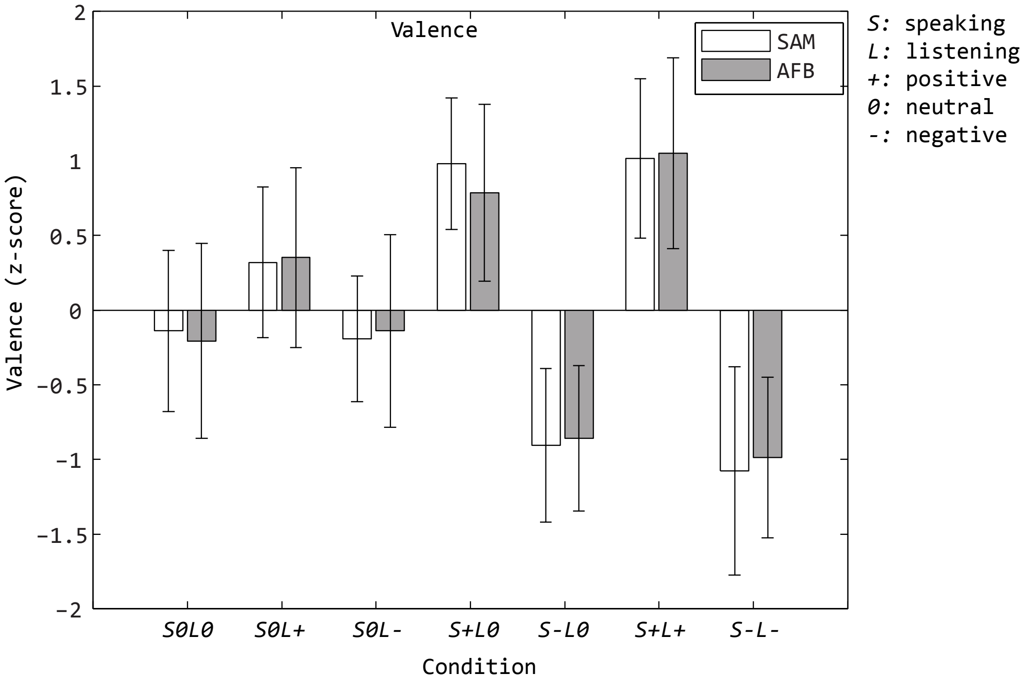


Figure 8: Means and standard deviations of SAM and AFB z-scores for the valence dimension for each of the seven experimental conditions.

## 4.1 Chinese versus non-Chinese

To test the effect of cultural background on the valence rating of the emotional expressions, a mixed MANOVA was conducted using condition as a within-subjects independent variable, cultural background as a between-subjects independent variable, and averaged valence score of both measurements as a dependent variable. The results showed no significant main effect for the cultural background on valence score $F(1,22) = 1.23$, $p = .64$, and no significant interaction between cultural background and condition $F(6,17) = 0.72$, $p = .28$.

Instead of looking for a difference between participants from different cultural backgrounds, the next step of the analysis focused on similarity in the ratings between these two groups. To examine the relationship between the ratings of Chinese and non-Chinese participants, we performed a correlation analysis based on the means for the seven conditions. The results showed that the scores on valence of the Chinese participants are significantly correlated with those of the non-Chinese participants $r(7) = .98$, $p < .001$. Although a difference in cultural background was expected, the result showed a high consistency in the evaluation of the emotional state between participants from different cultures. Hence, the results of the two groups of participants were grouped in the rest of the data analyses.

## 4.2 Positive versus Neutral versus Negative Emotional State

Participants were asked to rate seven conditions (i.e., different combinations of a positive, negative and neutral emotional state during the virtual human's speaking and listening phase). A repeated-measures MANOVA was conducted to study the effect of these conditions on averaged valence score of the SAM and AFB z-scores. The results showed a significant effect of condition on the valence rating, $F(6,18) = 59.50$, $p < .001$. Next, to run a priori comparisons, paired-sample *t*-tests were performed using the averaged valence scores of the SAM and AFB z-scores in all the conditions as paired variables. The results are shown in Table 1.

Table 1: Mean, SD and Mean difference of the valence rating of the different conditions.

| | | | Mean Difference / Conditions | | | | | |
|---|---|---|---|---|---|---|---|---|
| Condition | *M* | *SD* | *S0L+* | *S0L-* | *S+L0* | *S-L0* | *S+L+* | *S-L-* |
| *S0L0* | -0.17 | 0.50 | -0.51* | -0.008 | -1.06* | 0.71* | -1.21* | 0.86* |
| *S0L+* | 0.34 | 0.47 | | 0.50* | -0.55* | 1.22* | -0.70* | 1.37* |
| *S0L-* | -0.17 | 0.47 | | | -1.05* | 0.72* | -1.20* | 0.87* |
| *S+L0* | 0.88 | 0.39 | | | | 1.76* | -0.15 | 1.92* |
| *S-L0* | -0.88 | 0.42 | | | | | -1.91* | 0.15 |
| *S+L+* | 1.03 | 0.49 | | | | | | 2.06* |
| *S-L-* | -1.03 | 0.52 | | | | | | |

$H_0$: $\mu_1=\mu_2$, * $p < 0.05$.

To test whether the subjective valence score was correlated with the intensity weighted predicted valence values (see chapter 3.3 and Figure 5, and hereafter abbreviated as weighted valence values) for each condition, we calculated the Pearson correlation coefficient between the weighted valence values and the subjective scores averaged over the participants across the seven experimental conditions. This correlation was relatively high, *r*(7) = .93, *p* = .002. The following step in the analysis was to determine the deviation between the subjective valence scores and their corresponding expected valence value per experimental condition. To do so, we fitted a line through the three data points: *S+L+*, *S0L0* and *S-L-* using least-squares regression. Figure 9 shows this line, including the mean subjective valence scores of the remaining four conditions. Deviations of perceived valence from this line (for the lowly negative and positive videos) show to what extent the perceived valence is different from what is expected in case of an equal intensity in valence between the speaking and listening phase (noted as expected valence value in Figure 9).

One-sample *t*-tests revealed that when the virtual human showed neutral listening, both a positive (*S+L0*) and negative (*S-L0*) emotional expression during speaking had a more extreme valence than expected, i.e. the subjective score was more positive than the expected valence value in case of the positive emotional expression (*t*(23) = 2.69, *p* = .013) and more negative than the expected valence value in case of a negative emotional expression during speaking (*t*(23) = -6.14, *p* < .001). The opposite was seen for the impact of the listening phase. Considering the speaking phase with a neutral emotional expression, the subjective valence score for listening with a positive emotional expression (i.e., the *S0L+* condition) was significantly less positive than expected (*t*(23) = -3.08, *p* = .005). Similarly, the subjective valence score for listening with a negative emotional expression (i.e., the *S0L-* condition) was significantly less negative than the expected valence value (*t*(23) = 3.38, *p* = .003).

Moreover, the subjective valence score for the *S0L-* condition was almost equal (*t*(23) = 0.059, *p* = .95) to the subjective valence score for the *S0L0* condition (i.e., speaking with a neutral emotional expression and listening with a neutral emotional expression). Still, the subjective valence score of the *S0L+* condition (i.e., speaking with a neutral emotional expression and listening with a positive emotional expression) was significantly more positive than that for the *S0L0* condition ((*t*(23) = 2.92, *p* = .008). A direct comparison of the lowly positive or negative conditions provided a similar pattern. The subjective valence value for the *S-L0* condition (i.e., speaking with a negative emotional expression and listening with a neutral emotional expression) was significantly more negative than the subjective valence value for the *S0L-* condition (i.e., speaking with a neutral emotional expression and listening with a negative emotional expression), *t*(23) = 4.97, *p* < .001. Similarly, the subjective valance value for the *S+L0* condition (i.e., speaking with a positive emotional expression and listening with a neutral emotional expression) was significantly more positive than the subjective valence value for the *S0L+* condition (i.e., speaking with a neutral emotional expression and listening with a positive emotional expression), *t*(23) = 4.01, *p* = .001.

Together these observations imply that people do not perceive much difference between the virtual human showing neutral or negative listening behavior, but they do perceive a difference with the virtual human showing positive listening behavior. In conclusion, all these results support hypothesis 2, stating that the valence of the emotional expression during the listening phase of a conversation is perceived as less impactful compared to the emotional expression during the speaking phase.

Finally, we also compared the more extreme emotional conditions with the *S0L0* condition. The *S+L+* condition (*t*(23) = -9.00, *p* < .001) or *S-L-* condition (*t*(23) = 5.16, *p* < .001) with positive or negative emotional expressions in both the listening and speaking phase, respectively, strongly impact the perceived valence in the expected way.

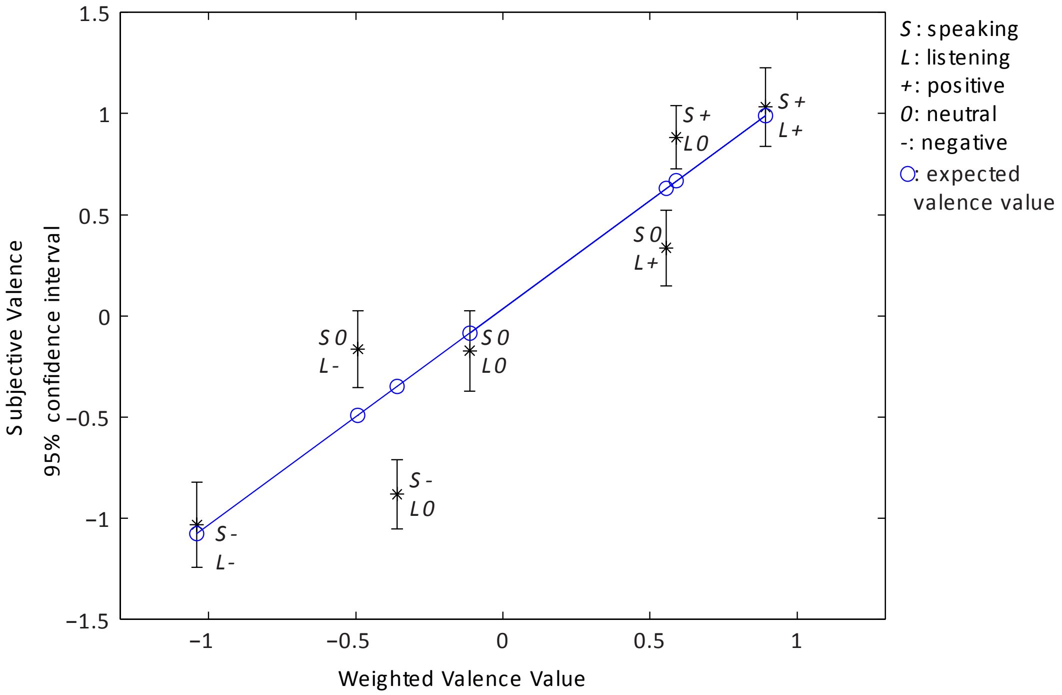


Figure 9: The relationship between intensity weighted predicted valence and the averaged subjective valence.

# 5. Discussion and conclusion

The experiment described in this paper is a human perception study on positive and negative emotions of a virtual human and how cultural background might affect the perception of these emotions. In a sense this study can be seen as a re-confirmation in virtual reality of what is known about human-human interaction in the actual world. Still this is an important validation step as conversations with virtual humans are increasingly used as part of gaming (e.g., Hudlicka and Delft (2009), training (e.g., Broekens, Harbers, et al. (2012), or psychotherapy (e.g., Opris et al. (2012).

The study found that both Chinese and non-Chinese participants could perceive the valence of the virtual human's emotional states and no significant difference between these two groups was found. Instead, the ratings of these two groups were highly correlated. The results show that the valence of the emotional states of the virtual human can be easily recognized by all participants independent of their cultural backgrounds. Hypothesis 1 is therefore not confirmed. On the contrary, our results support the idea of universality of the facial expression of emotion (Ekman, 1994; Matsumoto, 2007), and question the need for tailored made virtual reality applications which target different cultural groups or have multi-cultural users. Still, the results of this study may not be generally applicable to all cultures, since we here only evaluated possible differences in emotion perception between Chinese and non-Chinese people. Further studies are needed to extend our conclusion of universality of emotion perception of virtual human to people with other cultural backgrounds.

In addition, comparing the difference between conditions, it seems that the participants' perception of the valence was more influenced by the emotion of the virtual human while speaking than while listening; and so, this supports Hypothesis 2. Comparing the subjectively perceived valence scores with the expected valence values (Figure 9), the valence perceived by the participants in the conditions where the listening was neutral, but the speaking performed with a positive or negative emotional expression, was significantly more extreme than what was predicted from equal intensity between speaking and listening. Similarly, the perceived valence was less extreme than the weighted valence value when the speaking was neutral, but the listening performed with a positive or negative emotional expression. This shows the additional influence of verbal communication on valence recognition during a human-virtual human conversation. These findings seem to be in contrast to reports of Melo, Carnevale, and Gratch (2011), who claim that there is no difference in emotion perception between verbal and non-verbal communication. Their study however used text typing as verbal communication means between human and virtual human, which might explain the different finding. It seems not surprising that the combination of both verbal and non-verbal communication transfers more emotional information than the non-verbal communication only. Furthermore, the influence of the voice can be regarded as content independent because of the high consistency found between the Chinese and non-Chinese participants in this experiment. In other words, the results suggest that affective aspects can be conveyed in the speech even if the language is not understood.

The finding that the perceived valence of the emotion of the virtual human is more intense in the speaking phase than in the listening phase of a conversation may be extended with new research on how to control the level of emotion during these separate phases. Applications such as virtual reality exposure therapy for patients suffering from social phobia may be designed in a way to manipulate the potential phobic stressor using the virtual human’s emotional behavior. Further studies may exploit the difference in valence perception between the speaking and listening phase, and explore how to further optimize the persuasive power during these two phases, which may be beneficial for the design of many virtual applications involving human-virtual human conversation. Besides, this study only focuses on how individuals perceive the performance of a virtual human. It is also interesting to test the emotional influence on a human during a human-virtual human conversation. Whether the virtual human's emotion could lead or alter the content of the conversation could be an appealing topic in the persuasive computing area.

Two main conclusions may be drawn from the experiment, but there are also still a number of limitations. First, the virtual human only showed her upper body and no gestures were used to express emotion. However, in recent decades, more insights have become available on body expression (Gross, Crane, & Fredrickson, 2010; Kleinsmith & Bianchi-Berthouze, 2013). It would therefore be interesting to examine how our findings would be affected when the virtual human used its full-body to express emotions. Second, the position of the virtual human was fixed in the current study. It would be interesting to test the emotional impact of manipulating the virtual human's position, for example, far away versus nearby (Broekens, Qu, et al., 2012). Third, the face model of the virtual human we used in this study was generated by FaceGen with the ethnicity parameter set at Southeast Asia. However, no empirical validation was done to confirm the ethnic appearance of the virtual human. Fourth, the study described in this paper only focused on the valence dimension of the emotion, neglecting so far the other two dimensions of emotion, namely arousal and dominance. Including the additional two dimensions would allow to study more complex emotions, for example, fear, surprise, etc. Despite of the limitations, the results of this paper suggest a superior impact on perceiving the virtual human's emotional state during its speaking phase, and a potential independence of the perceived valence of the virtual human's emotion with cultural background. These findings could help designers to focus their attention upon creating and evaluating virtual human with appropriate emotional expressions, which may help to improve the overall experience of virtual environments.

# Acknowledgements

This study is supported in part by the Chinese Scholarship Council (No.2008609199) and the COMMIT project - Interaction for Universal Access.

# Reference

Bailenson, J. N., Yee, N., Merget, D., & Schroeder, R. (2006). The Effect of Behavioral Realism and Form Realism of Real-Time Avatar Faces on Verbal Disclosure, Nonverbal Disclosure, Emotion Recognition, and Copresence in Dyadic Interaction. *Presence: Teleoperators and Virtual Environments, 15*(4), 359-372.

Barrett, L. F. (2006). Solving the emotion paradox: categorization and the experience of emotion. *Personality and social psychology review : an official journal of the Society for Personality and Social Psychology, Inc, 10*, 20-46.

doi: 10.1207/s15327957pspr1001_2

Bartneck, C. (2001). Affective expressions of machines. *CHI '01 extended abstracts on Human factors in computing systems*, 189-190.

Bates, J. (1994). The role of emotion in believable agents. *Communications of the ACM, 37*(7), 122-125.

Bradley, M. M., & Lang, P. J. (2007). The International Affective Digitized Sounds (IADS-2): Affective ratings of sounds and instruction manual. *Emotion*, 29-46.

Brand, M. (1999). Voice puppetry. *Proceedings of the 26th annual conference on Computer graphics and interactive techniques SIGGRAPH 99*, 21-28.

Bregler, C. (1997). Video rewrite: Driving visual speech with audio. *Proceedings of SIGGRAPH '97*, 1-8.

Brinkman, W.-P., Hartanto, D., Kang, N., de Vliegher, D., Kampmann, I. L., Morina, N., . . . Neerincx, M. A. (2012). *A virtual reality dialogue system for the treatment of social phobia.* Paper presented at the CHI'12 extended abstracts on human factors in computing systems.

Brinkman, W.-P., Veling, W., Dorrestijn, E., Sandino, G., Vakili, V., & van der Gaag, M. (2011). Virtual reality to study responses to social environmental stressors in individuals with and without psychosis. *Studies in Health Technology and Informatics, 167*, 86-91.

Broekens, J., & Brinkman, W.-P. (2009). Affectbutton: Towards a standard for dynamic affective user feedback. *Affective Computing and Intelligent Interaction and Workshops, 2009*, 1-8.

Broekens, J., & Brinkman, W.-p. (2013). AffectButton: A method for reliable and valid affective self-report. *International Journal of Human-Computer Studies, 71*(6), 641 - 667.

Broekens, J., Harbers, M., Brinkman, W.-P., Jonker, C., Van den Bosch, K., & Meyer, J.-J. (2012). Virtual reality negotiation training increases negotiation knowledge and skill. *Intelligent Virtual Agents*, 218-230.

Broekens, J., Harbers, M., Brinkman, W.-P., Jonker, C., Van den Bosch, K., & Meyer, J. J. (2011). Validity of a Virtual Negotiation Training. *Intelligent Virtual Agents*, 435-436.

Broekens, J., Pronker, A., & Neuteboom, M. (2010). Real time labeling of affect in music using the affectbutton. *Proceedings of the 3rd international workshop on Affective interaction in natural environments*, 21-26.

Broekens, J., Qu, C., & Brinkman, W.-P. (2012). Dynamic Facial Expression of Emotion Made Easy. *Technical report* (pp. 1-30): Interactive Intelligence, Delft University of Technology.

Cassell, J., Pelachaud, C., Badler, N., Steedman, M., Achorn, B., Becket, T., . . . Stone, M. (1994). Animated Conversation: Rule-based Generation of Facial Expression, Gesture and Spoken Intonation for Multiple Conversational Agents. *Proc. of ACM SIGGRAPH*, 413-420.

Cassell, J., & Thorisson, K. R. (1999). The Power of a Nod and a Glance: Envelope vs. Emotional Feedback in Animated Conversational Agents. *Applied Artificial Intelligence, 13*, 519-538.

Cerezo, E., & Baldassarri, S. (2008). Affective Embodied Conversational Agents for Natural Interaction. *Affective Computing*, 329-354.

Chuang, E., & Bregler, C. (2002). Performance driven facial animation using blendshape interpolation. *Computer Science Technical Report, Stanford University*.

Church, T., & Katigbak, M. (1998). Language and organisation of Filipino emotion concepts: Comparing emotion concepts and dimensions across cultures. *Cognition & Emotion, 12*(1), 63-92.

Cohen, M. M., & Massaro, D. W. (1993). Modeling coarticulation in synthetic visual speech. *Models and Techniques in Computer Animation*, 139-156.

Core, M., Traum, D., Lane, H., & Swartout, W. (2006). Teaching negotiation skills through practice and reflection with virtual humans. *Simulation*.

Cowell, A. J., & Stanney, K. M. (2003). Embodiment and Interaction Guidelines for Designing Credible, Trustworthy Embodied Conversational Agents. *4th International Workshop on Intelligent Virtual Agents IVA 2003, 2792*, 301-309.

Darwin, C. (1872). The Expression of Emotion in Man and Animals. *New York: Philosophical Library*.

Dotsch, R., & Wigboldus, D. H. J. (2008). Virtual prejudice. *Journal of Experimental Social Psychology, 44*(4), 1194-1198.

Ekman, P. (1994). Strong evidence for universals in facial expressions: a reply to Russell's mistaken critique. *Psychological Bull, 115*(2), 268-287.

Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of personality and social, 17*(2), 124-129.

Ekman, P., Friesen, W. V., & Hager, J. C. (2002). Facial Action Coding System. *A Human Face, 97*, 4-5.

Ekman, P., Friesen, W. V., O'Sullivan, M., Chan, A., Diacoyanni-Tarlatzis, I., Heider, K., . . . Ricci-Bitti, P. (1987). Universals and cultural differences in the judgments of facial expressions of emotion. *J Pers Soc Psychol, 53*(4), 712-717.

Ekman, P., Rolls, E. T., Perrett, D. I., & Ellis, H. D. (1992). Facial expressions of emotion: An old controversy and new findings. *Philosophical Transactions: Biological Sciences, 335*, 63-69.

Elfenbein, H. A. (2003). Universals and cultural differences in recognizing emotions. *Current Directions in Psychological*, 159-164.

Elfenbein, H. A., & Ambady, N. (2002). Is there an in-group advantage in emotion recognition? *Psychological Bulletin, 128*(2), 243-249.

Elfenbein, H. A., Beaupre, M., Levesque, M., & Hess, U. (2007). Toward a dialect theory: cultural differences in the expression and recognition of posed facial expressions. *Emotion (Washington, D.C.), 7*(1), 131-146.

Endrass, B., Rehm, M., & Lipi, A. (2011). Culture-related differences in aspects of behavior for virtual characters across Germany and Japan. *Proceedings of AAMAS'11, 2*, 441-448.

Ersotelos, N., & Dong, F. (2008). Building highly realistic facial modeling and animation: a survey. *The Visual Computer, 24*(1), 13-30.

Ezzat, T., Geiger, G., & Poggio, T. (2004). Trainable videorealistic speech animation. *Sixth IEEE International Conference on Automatic Face and Gesture Recognition 2004 Proceedings*, 57-64.

Fox, E. (2008). Emotion Science Cognitive and Neuroscientific Approaches to Understanding Human Emotions. *Palgrave Macmillan*.

Gratch, J., Rickel, J., Andre, E., Cassell, J., Petajan, E., Badler, N. I., & Jeff, R. (2002). Creating interactive virtual humans: Some assembly required. *Intelligent Systems, IEEE, 17*(4), 54-63.

Gross, M. M., Crane, E. a., & Fredrickson, B. L. (2010). Methodology for Assessing Bodily Expression of Emotion. *Journal of Nonverbal Behavior, 34*, 223-248. doi: 10.1007/s10919-010-0094-x

Haring, M., Bee, N., & Andre, E. (2011). Creation and Evaluation of emotion expression with body movement, sound and eye color for humanoid robots. *RO-MAN, 2011 IEEE*, 204-209.

Hofstede, G. (2001). Culture's consequences: Comparing values, behaviors, institutions and organisations across nations. *Sage Publications, Thousand Oaks*.

Hudlicka, E., & Delft, T. U. (2009). Foundations for modelling emotions in game characters: Modelling emotion effects on cognition. *Affective Computing and Intelligent Interaction and Workshops, ACII 2009*.

Irtel, H. (2007). PXLab: The Psychological Experiments Laboratory [online]. *Mannheim (Germany): University of Mannheim.*

Isbister, K. (2006). Better Game Characters by Design: A Psychological Approach. *Education*.

Jack, R. E., Garrod, O. G. B., Yu, H., Caldara, R., & Schyns, P. G. (2012). Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences, 109*(19), 7241-7244.

Jan, D., Herrera, D., & Martinovski, B. (2007). A computational model of culture-specific conversational behavior. *IVA '07 Proceedings of the 7th international conference on Intelligent Virtual Agents*, 45-56.

Kahler, K., Haber, J., & Seidel, H.-P. (2001). Geometry-based Muscle Modeling for Facial Animation. *Proc. of Graphics Interface*, 37-46.

Keltner, D., & Ekman, P. (2000). Facial expression of emotion. *Handbook of Emotions, 2nd Edition*, 236-249.

Kleinsmith, A., & Bianchi-Berthouze, N. (2013). Affective Body Expression Perception and Recognition: A Survey. *IEEE Transactions on Affective Computing, 4*, 15-33. doi: 10.1109/T-AFFC.2012.16

Kleinsmith, A., De Silva, P. R., & Bianchi-Berthouze, N. (2006). Cross-cultural differences in recognizing affect from body posture. *Interacting with Computers, 18*(6), 1371-1389.

Kulms, P., Kramer, N. C., Gratch, J., & Kang, S.-H. (2011). It's in Their Eyes: A Study on Female and Male Virtual Humans' Gaze. *Intelligent Virtual Agents*, 80-92.

Lance, B. J., Rey, M. D., & Marsella, S. C. (2008). A model of gaze for the purpose of emotional expression in virtual embodied agents. *AAMAS '08 Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems, 1*, 12-16.

Lang, P. J. (1995). The emotion probe. Studies of motivation and attention. *American Psychologist, 50*(5), 372-385.

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (1999). International affective picture system (IAPS): Technical manual and affective ratings. *Psychology*.

Lang, P. J., Bradley, M. M., & Cuthbert, B. N. (2008). International affective picture system (IAPS): Affective ratings of pictures and instruction manual. *Technical Report A-8*.

Lee, J., & Marsella, S. (2012). Modeling Speaker Behavior: A Comparison of Two Approaches. *Intelligent Virtual Agents*, 161-174.

Lee, J., Prendinger, H., Neviarouskaya, A., & Marsella, S. (2009). Learning models of speaker head nods with affective information. *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, 1-6.

Lee, Y., Terzopoulos, D., & Walters, K. (1995). Realistic modeling for facial animation. *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques SIGGRAPH 95*, 55-62.

Link, M., Armsby, P., Hubal, R. C., & Guinn, C. I. (2006). Accessibility and acceptance of responsive virtual human technology as a survey interviewer training tool. *Computers in Human Behavior, 22*, 412-426. doi: 10.1016/j.chb.2004.09.008

Litwinowicz, P., & Williams, L. (1994). Animating images with drawings. *Proceedings of the 21st annual conference on Computer graphics and interactive techniques SIGGRAPH 94*, 409-412.

MacDorman, K. F. K., Coram, J. J. A., Ho, C.-C. C., & Patel, H. (2010). Gender differences in the impact of presentational factors in human character animation on decisions in ethical dilemmas. *Presence: Teleoperators and Virtual Environments, 19*(3), 213-229.

Matsumoto, D. (2002). Methodological requirements to test a possible in-group advantage in judging emotions across cultures: Comment on Elfenbein and Ambady (2002) and evidence. *Psychological Bulletin, 128*(2), 236-242.

Matsumoto, D. (2007). Emotion judgments do not differ as a function of perceived nationality. *International Journal of Psychology, 42*(3), 207-214.

Mehrabian, A., & Russell, J. A. (1974). An Approach to Environmental Psychology. *MIT Press, Cambridge, MA, USA; London, UK*.

Melo, C. d., Carnevale, P., & Gratch, J. (2011). The effect of expression of anger and happiness in computer agents on negotiations with humans. *The Tenth International Conference on Autonomous Agents and Multiagent Systems*, 2-6.

Moridis, C. N., & Economides, A. A. (2012). Affective Learning: Empathetic Agents with Emotional Facial and Tone of Voice Expressions. *IEEE Transactions on Affective Computing, 3*(3), 260-272.

Morris, J. D. (1995). Observations : SAM The Self-Assessment Manikin An Efficient Cross-Cultural Measurement Of

Emotional Response. *Journal of Advertising Research, 35*(6), 63-68.

Ochsman, R. B., & Chapanis, A. (1974). The effects of 10 communication modes on the behavior of teams during co-operative problem-solving. *International Journal of ManMachine Studies, 6*(5), 579-619.

Opris, D., Pintea, S., Garcia-Palacios, A., Botella, C. M., Szamoskozi, S., & David, D. (2012). Virtual reality exposure therapy in anxiety disorders: a quantitative meta-analysis. *Depression and anxiety, 29*, 85-93. doi: 10.1002/da.20910

Parke, F. I. (1972). Computer generated animation of faces. *Proceedings of the ACM annual conference, 1*, 451-457.

Parke, F. I. (1974). A parametric model for human faces. *The University of Utah, Doctoral Dissertation*.

Petrushin, V. (1999). Emotion in speech: Recognition and application to call centers. *Artificial Neu. Net. In Engr.(ANNIE'99)*, 7-10.

Picard, R. W. (1998). Toward agents that recognize emotion. *Actes Proceedings IMAGINA*, 153-165.

Picard, R. W., Vyzas, E., & Healey, J. (2001). Toward machine emotional intelligence: Analysis of affective physiological state. *Pattern Analysis and Machine Intelligence, IEEE Transactions on, 23*(10), 1175-1191.

Platt, S. M., & Badler, N. I. (1981). Animating facial expressions. *ACM SIGGRAPH Computer Graphics, 15*(3), 245-252.

Qiu, L., & Benbasat, I. (2005). Online consumer trust and live help interfaces: The effects of text-to-speech voice and three-dimensional avatars. *International Journal of Human-Computer Interaction, 19*, 37-41.

Reeves, B., & Nass, C. (1996). The Media Equation. *Cambridge University Press*.

Rizzo, A., Lange, B., Buckwalter, J., Forbell, E., Kim, J., Sagae, K., . . . Kenny, P. (2011). An intelligent virtual human system for providing healthcare information and support. *Stud Health Technol Inform, 163*, 503-509.

Russell, J. A. (1991). Culture and the categorization of emotions. *Psychological bulletin, 110*, 426-450.

Ruttkay, Z., & Pelachaud, C. (2005). *From Brows to Trust: Evaluating Embodied Conversational Agents*: Springer.

Scherer, K. R. (1995). Expression of emotion in voice and music. *Journal of voice : official journal of the Voice Foundation, 9*(3), 235-248.

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication, 40*, 227-256.

Schiano, D. J., Ehrlich, S. M., Rahardja, K., & Sheridan, K. (2000). Face to interface: facial affect in (hu)man and machine. *Proceedings of ACM CHI 2000*, 193-200.

Schroder, M. (2004). Speech and Emotion Research: an overview of research frameworks and a dimensional approach to emotional speech synthesis. *Research Report of the Institute of Phonetics*.

Slater, M., Pertaub, D.-P., & Steed, A. (1999). Public speaking in virtual reality: facing an audience of avatars. *IEEE Computer Graphics and Applications, 19*(2), 6-9.

Tomkins, S. S. (1962). Affect, Imagery, Consciousness: Vol 1. The Positive Affects. *New York: Springer*.

Tomkins, S. S. (1963). Affect, Imagery, Consciousness: Vol 2. The Negative Affects. *New York: Springer*.

Tsapatsoulis, N., Raouzaiou, A., Kollias, S., Cowie, R., & Douglas-Cowie, E. (2002). Emotion Recognition and Synthesis Based on MPEG-⍰4 FAPs. *MPEG-4 facial animation the standard implementations applications*.

Waters, K. (1987). A Muscle Model for Animating Three-Dimensional Facial Expression. *Comput Graph SIGGRAPH Proc, 21*(4), 17-24.

Wierzbicka, A. (1995). Emotions across languages and cultures: Diversity and universals. *Cambridge University Press, Cambridge*.

Wong, J., & McGee, K. (2012). Frown More, Talk More: Effects of Facial Expressions in Establishing Conversational Rapport with Virtual Agents. *Intelligent Virtual Agents*, 419-425.

Yun, C., Deng, Z., & Hiscock, M. (2009). Can local avatars satisfy a global audience? A case study of high-fidelity 3D facial avatar animation in subject identification and emotion perception by US and international groups. *Computers in Entertainment, 7*(2), 1-25.

Zhang, Q. Z. Q., Liu, Z., Quo, G. Q. G., Terzopoulos, D., & Shum, H.-Y. S. H.-Y. (2006). Geometry-driven photorealistic facial expression synthesis. *IEEE Transactions on Visualization and Computer Graphics, 12*(1), 48-60.