Moosaei, M. and Riek, L.D. (2013). "Evaluating Facial Expression Synthesis on Robots". In Proceedings of the HRI Workshop on Applications for Emotional Robots at the 8th ACM International Conference on Human-Robot Interaction (HRI). Tokyo, March 3, 201

# Evaluating Facial Expression Synthesis on Robots

Maryam Moosaei and Laurel D. Riek
Department of Computer Science and Engineering
University of Notre Dame
Notre Dame, IN, 46556, USA
{mmoosaei, lriek}@nd.edu

*Abstract*—**In this paper, we outline techniques for evaluating facial expression synthesis methods on robots and virtual avatars. We describe common evaluative methods, including subjective and computationally-based techniques. We also present our new evaluation model which combines advantages of both. Finally, we discuss challenges in performing synthesis evaluation that are unique to the HRI research community.**

## I. Introduction

Robotics is a rapidly growing industry. Social robots especially are predicted to become more common in our daily lives [1]. These robots include socially and physically assistive robots, domestic care robots, toys, coaches, user interfaces to smart homes, and household robots. Robots with human-like faces may be more readily perceived as pleasant and user-friendly, and may allow humans to more easily predict their intentions. Thus, as robots begin to have a strong presence in our everyday lives, there is a strong desire among researchers and robot designers in the social robotics community to enable them to have natural face-to-face interaction with humans. Synthesizing facial expressions may increase the perceived naturalness of human-robot interaction (HRI) [2]. Research in this area can have several applications beyond HRI including: animation, teleconferencing, facial surgery, computer supportive collaborative work, intelligent tutoring, and many areas of healthcare.

There are many examples in the literature proposing techniques for performing facial expression synthesis on robots (c.f., [3] [4] [5] [6]); however, less attention has been paid toward developing a standard evaluation method to compare different synthesis systems. Currently, the two most common ways to evaluate facial expression synthesis is to perform a subjective or computationally-based evaluation. In the first one, subjects judge the synthesized expressions, and in the second one each synthesized expression is quantitatively compared with a predefined computer model. In this paper, we investigate the strengths and weaknesses of these methods, as well as propose a new model that combines the advantages of both.

## II. Background

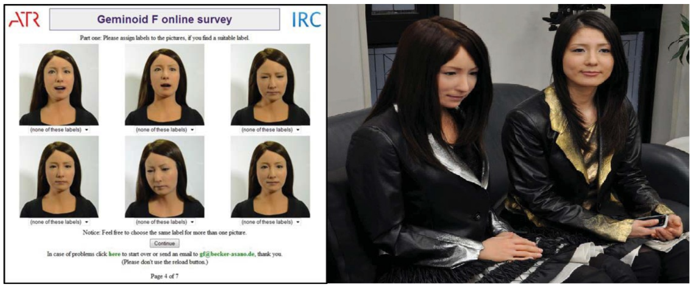
In the broader field of human-computer interaction, several researchers have studied the classification of evaluation methods. For example, Benoit et al. [7] proposed that based on the evaluation goal there are three types of evaluations for a multi-modal synthesis system:

- Adequacy evaluation: evaluates how well the system provides requirements determined by user's needs. For example, a consumer report is a kind of adequacy evaluation.
- Diagnostic evaluation: evaluates system performance against a taxonomy, considering possible usage in the future. System developers usually use this evaluation.
- Performance evaluations: evaluates the system in a specific area. To do so, a well-defined performance baseline is required for comparison. For example, an older version of the same system or a different system which supports the same functionality.

Of the three techniques, performance evaluation is the most common in the literature [7]. Performance evaluation for a system requires defining the characteristics one is interested in evaluating (e.g. accuracy, error rate, or recognition rate) as well as an appropriate method for measuring them. Some criteria for performance evaluation of facial expression synthesis systems can be task completion time, complexity, naturalness of expressions, static versus dynamic expressions, and recognition rate. Generally, there are three ways of performing a performance evaluation for a synthesis technique [8]:

- User-based: This evaluation method involves users completing some predefined tasks which match the goals the system is designed for. By analyzing data collected from users, the system performance is evaluated.
- Expert-based: This evaluation method is similar to user-based except users are expert people in that area of synthesis. For example, in case of facial expression synthesis users may have expertise in decoding facial expressions (e.g. certified Facial Action Coding System (FACS) coders).
- Theory-based: This evaluation method does not require live human-machine interaction. Instead, it tries to predict the user's behavior or performance based on a theoretically derived model.

From another point of view, evaluative methods can be divided into two classes: quantitative and qualitative [9]. Quantitative evaluations compare synthesized values with real values to investigate whether synthesized movements are correct or not [7]. For example, one can compare the values of lip height and lip width in a synthetic smile with



Geminoid F online survey

Part one: Please assign labels to the pictures, if you find a suitable label.

Notice: Feel free to choose the same label for more than one picture.

In case of problems click here to start over or send an email to gf@becker-asano.de, thank you.
(Please don't use the reload button.)

Page 4 of 7

Fig. 1. Subjective evaluation example from Becker-Asano and Ishiguro [17]; left: the interface, right: Geminoid F (left) and its model person (right).

corresponding values of a human subject. By using an image-based FACS analysis, one can compare muscle contradiction in a real image with a synthetic expression. However, this is a complicated task since different facial parameters may have different levels of importance in conveying an emotion and one needs to find an appropriate weight for each of them [7]. For example, Bassili [10] and Gouta and Miyamoto [11] found that negative emotions are usually expressed by the upper part of the face while positive emotions are expressed in the lower part [12]. Computing appropriate weightings of each facial part in an emotional expression is still an open question [13].

A qualitative evaluation compares the intelligibility of a synthesized expression with the intelligibility of the same emotion expressed by a human. In other words, a qualitative evaluation checks how the synthesized expressions are perceived [7].

Berry et al. [14] proposed that evaluations should be done on different levels of a synthesis system including:

- Micro level: observes just one aspect of the robot separately of other parts (e.g. just lip movements).
- User level: evaluates the reaction of users to the system.
- Application level: evaluates the system within a specific application.

The most common evaluation techniques used by researchers in the field of facial expression synthesis are subjective evaluation, expert evaluation, and computational evaluation.

### A. *Subjective Evaluation*

In the literature, subjective evaluation is the most commonly used approach to evaluate facial expression synthesis systems. Subjects observe synthesized expressions and then answer a predefined questionnaire. By analyzing collected answers, researchers evaluate the expressions of their robot or virtual avatar. Although user studies are costly and time consuming, they provide valuable information about the acceptability and believability of a robot. Moreover, humans are very sensitive to errors in facial movements. A wrong movement, a wrong duration, or a sudden transition between facial expressions can be detected by a human. This is especially true when the robot or virtual agent is more realistic in appearance [15].

A subjective evaluation requires a well-defined experimental procedure and a careful analysis of the results. Several methodological issues should be considered when choosing participants, such as: their average age, their educational level, their cultural background, their native language, and the gender ratio of participant. A second experimental design concern is that some facial expressions have different meanings in different cultures; for example, in India, the same head nod used to express agreement may convey disagreement in a Western country [16]. Becker-Asano and Ishiguro, with robot Geminoid F, studied the effects of intercultural differences in percieving the robots' facial expressions [17]. They found that Asian participants showed low agreement in what constituted a fearful or surprised expression in contrast to American and European participants.

A third experimental design concern is how to best select subjects to participate in the evaluation. Christoph [18] did research in this area, suggesting that an ideal group of subjects are those most likely to use the robot in the future. For example, a robot designed for autism therapy should be evaluated by autistic children since the robot is intended to be useful for them.

Another important aspect of subjective evaluation experimental design is adequately preparing subjects for interaction with the robot. For example, Riek et al. [19] showed subjects a picture of an unusual-looking zoomprhic robot before the experiment in order to help them adequately habituate to it and avoid uncanny effects. This step may prove particularly important for subjective evaluation of very realistic looking robots so as not to shock participants.

Several examples of strong experimental design for subjective synthesis evaluation are in the literature. For example, Becker-Asano and Ishiguro [17] performed a subjective study to compare expressivity of Geminoid F's five basic facial expressions with the expressions of the real model person. After watching each image of the robot expressing different facial expressions, subjects could choose one of the six labels including *angry*, *fearful*, *happy*, *sad*, *surprise* or *neutral*. To study intercultural differences in recognition accuracy, they performed the experiment in German, Japanse and English language. The interface the researchers used in their subjective evaluation as well as the model person can be seen in Figure 1.

In another study, Mazzei et al. used a subjective method to evaluate the robotic face FACE in conveying emotion to children with Autism Spectrum Disorders (ASD) [3]. They recruited participants from both the ASD and non-ASD

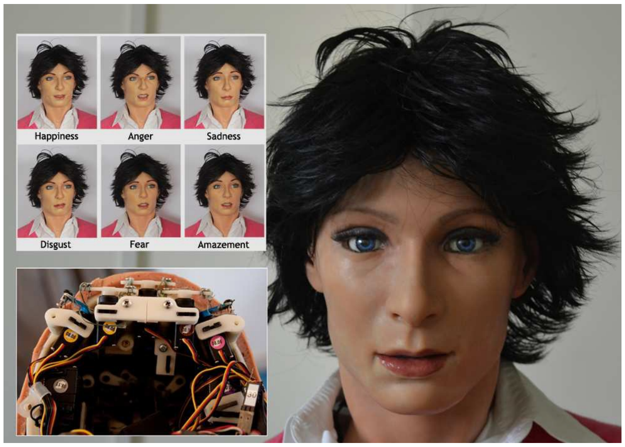


Fig. 2. Subjective evaluation example from Mazzei et al. [3].

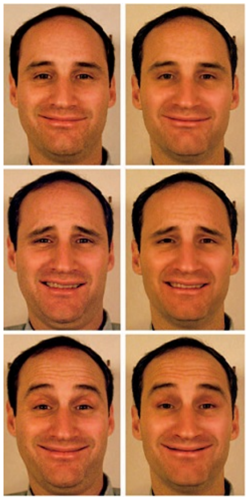


Fig. 3. An example of a side-by-side comparison. On the left is the real image, and on the right is the synthesized [21].

population. Participants were asked to recognize and label the emotions expressed by the robot including *happiness*, *anger*, *sadness*, *disgust*, *fear* and *surprise*. The correctness of the labels were determined by a therapist. The interface they used for their subjective experiment is depicted in Figure 2.

Another kind of subjective evaluation is a side-by-side comparison or “copy synthesis” in which researchers try to synthesize facial expressions which best match a recorded video or image [20][21]. Subjective evaluation about the quality of synthesis is done by visual comparison between the original videos and the synthesized ones. A sample figure of this side-by-side comparison is shown in Figure 3.

Subjective evaluation is also very common for analyzing human perception of virtual agents [14][22]. However, in the case of physical robots, their physicality changes both the physical generation of synthesis as well as the evaluation. Moving motors on a robot is different and more complicated than animating a virtual character. The number of motors, their speed, their range of motion, and the synchronization between them are among the factors which make the evaluation more complicated. Several evaluation metrics should be defined. We are concerned about questions such as: does motor noise affect immersion by being distracting? Are we more sensitive to perfection, speed or realism in physical motion versus virtual motion?

### B. *Expert Evaluation*

In an expert evaluation, a trained person, usually a psychologist, evaluates the synthesized facial expressions to determine whether the system provides predefined criteria or not. However, recruiting a sufficient number of experts is not easy. Moreover, often the robot is ultimately designed for users from the general population, not highly trained experts. A sole reliance on experts could create acceptability problems of the robot down the road.

### C. *Computational Evaluation*

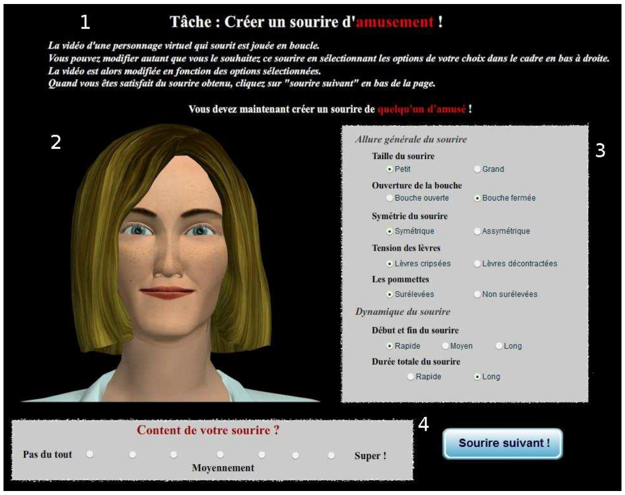
A computational evaluation aims to provide a fast and fully automatic evaluation method which does not require human judges. Designing such an evaluation can remove many of the aforementioned problems with subjective evaluations. This evaluation method requires one to develop an accurate computer model of facial parts for each facial expression. Based on this model, the system compares the synthesized model with the computer model.

Some researchers have worked on developing a computational model of facial parts such as muscular and skin models or lip shapes [12][23][24]. However, designing an accurate model for each facial part is very complicated. Moreover, one cannot use the same model for different robot platforms because each robot has its own characteristics such as the number of control points. For example, the physical robot Doldori does not have any control points for cheek movements [25].

## III. Proposed method based on combining subjective and computational evaluation methods

An ideal evaluation method should incorporate the advantages of both subjective and computational evaluations. Thus, we are developing such a method. During our development, we consider two important facts. First, social robots are designed to interact with people. Therefore, the robot should be visually appealing for people and it is important to investigate how people perceive the robot by involving them in the evaluation. Second, different synthesis methods should be compared on the same robot platform in order to have a fair evaluation.

Our model is based on Ochs et al. [26]. They developed the “E-smiles-creator” as a web application to analyze facial parameters of a virtual character for different kinds of smiles (*amused*, *polite*, and *embarrassed*). In E-smiles-creator the



Tâche : Créer un sourire d'amusement !

*La vidéo d'une personnage virtuel qui sourit est jouée en boucle.*
*Vous pouvez modifier autant que vous le souhaitez ce sourire en sélectionnant les options de votre choix dans le cadre en bas à droite.*
*La vidéo est alors modifiée en fonction des options sélectionnées.*
*Quand vous êtes satisfait du sourire obtenu, cliquez sur "sourire suivant" en bas de la page.*

Vous devez maintenant créer un sourire de quelqu'un d'amusé !

*Allure générale du sourire*

Taille du sourire
● Petit ○ Grand

Ouverture de la bouche
○ Bouche ouverte ● Bouche fermée

Symétrie du sourire
● Symétrique ○ Assymétrique

Tension des lèvres
● Lèvres cripsées ○ Lèvres décontractées

Les pommettes
● Surélevées ○ Non surélevées

*Dynamique du sourire*

Début et fin du sourire
● Rapide ○ Moyen ○ Long

Durée totale du sourire
○ Rapide ● Long

Content de votre sourire ?
Pas du tout ○ ○ ○ ○ ○ ○ ○ Super !
Moyennement

Fig. 4. E-smiles-creator [27].

user is asked to create different kinds of smiles on the face of a virtual agent by manipulating its control points. Subjects also choose their degree of satisfaction at the end.

The interface designed by Ochs et al. [26] is shown in Figure 4. In this interface, the kind of smile that the user should produce (e.g. *polite smile*) is in the upper quadrant and all possible parameters that the user can change (e.g. duration) are shown in left quadrant. Each time a user changes one parameter, the new video of the expression is played and the users also choose their degree of satisfaction. Considering all possible parameters of their virtual agent, one may produce as many as 192 different kinds of smiles. Then by using a decision tree they characterized parameters for each kind of smile. In the decision tree, they accounted for the degree of satisfaction for each user-produced smile. Therefore those with a higher degree of satisfaction are counted as more important.

We use this technique in the subjective part of our evaluation model. Our evaluation method has three steps:

*Step I:* We need a baseline to compare all the methods against. This baseline should be the best expression that a human can perceive from that robot with regards to all characteristics and limitations of the robot. In our proposed evaluation model this baseline is determined with a subjective study. This step requires designing appropriate software as an interface between the participant and the robot. Mazzei et al. designed such an interface to control the motors of the physical robot FACE by subjects [3]. In this study, subjects are asked to change control points of the robot to create a list of different expressions. For each expression they are asked to create the expression such that they think is the best expression they can create with that robot. At the end, they also choose their degree of satisfaction on a predefined scale.

*Step II:* In the second step, a decision tree is used to analyze the parameters which subjects chose for each expression. For each of them we also consider a weighting based on satisfaction degree. Therefore, expressions created with a higher degree of satisfaction will have greater effect.

*Step III:* The third step is comparing the synthesis method we want to evaluate with the model in the second step. This part is done computationally by comparing each parameter in the synthesis output with its corresponding parameter from step II. In other words, we want to compare the synthesized expression computationally with a baseline we produced in step one by a subjective study. Step three has the same procedure as a computationally-based evaluation in Section II.C, except instead of having a computer model of expressions, we develop the baseline from a user-based study.

Our proposed evaluation model can be used for evaluating a synthesis method on both the micro level and application level [14]. For the micro level, one can perform all three steps of our model on just one aspect of the robot face (e.g. just lip movements). For the application level, one can perform all three steps of our model within a specific application (e.g. autism therapy). However, for a user level evaluation, traditional subjective based methods should be employed since the focus of this evaluation is solely on people's reaction to the robot.

## IV. Challenges in Synthesis Evaluation

In evaluating synthesized facial expressions, realism, naturalness, pleasantness, believability, and acceptability of expressions are important. Thus, evaluation of facial expression synthesis for both robots and virtual avatars is challenging. One reason is that defining the evaluation criteria is frequently challenging since qualitative aspects play an important role [7]. Comparing different synthesis methods is difficult without having a standard evaluation criteria.

There are several other issues to be considered in evaluating facial expressions. Having an aesthetically appealing robotic face is not enough to have it believable [28][29]. To synthesize realistic and believable expressions robotic faces should be consistent, have a smooth transition, and match the context, state of mind and perceived personality of the robot [29]. Facial expressions of the robot should be synchronized with other behaviors such as head and body movements.

Modeling all the above parameters for a computational evaluation method is complicated. Additionally, computational evaluation requires determining a computer model for different facial expressions (e.g. *neutral face*, *happy*, and *sad*) with different intensities as a baseline to compare different expressions with. Defining such models is challenging.

Therefore, having a computational evaluation method without any human which works with all robotic platforms and all different synthesis methods is still an open question.

The appearance of the robot may affect judgment about synthesized expressions. Kidd et al. [30] performed a subjective study with a physical robot and an animated character and found that a physical robot is perceived more informative for human subjects than a virtual character. Numerous characteristics may affect a person's judgement, such as whether the robot is physically co-located with the participant, the robot's morphology, its perceived gender, and its perceived age. Besides, choosing the appropriate synthesis method greatly depends on the robot one wants to implement synthesis on. The number of motors of a physical robot or control points of a virtual avatar has a great effect on choosing a synthesis method.

To evaluate a synthesis system, synthesis quality may be measured computationally as well as by human judgment. Related to human evaluation, several psychological issues should be considered such as person's underlying attitude toward robots, as well as individual differences which may affect their judgment [31][32]. Also there are always human errors in subjective studies. For example separating fear from disgust is hard for human subjects especially, on facial robots without nose wrinkles [25][33].

From this set of challenges one can see the diversity of factors which should be considered when selecting the best synthesis method. It is necessary to carefully evaluate these choices and their effect on the robot. Evaluation of synthesis systems is still relatively new and no benchmarks or standard evaluation procedures exist in the literature [14].

## V. Conclusion

For social robots to become acceptable in our daily lives, having natural human robot interaction is important. Therefore, it is important to make robots capable of generating appropriate expressions understandable by humans. Developing a standard method for evaluation of synthesized behavior is challenging. Comparing different systems with varying characteristics and parameters is relatively unexplored in the literature. In this paper, we presented techniques for facial expression synthesis evaluation and outline major challenges in this field in order to guide researchers. We also proposed a model for evaluating facial expression synthesis methods which incorporates both subjective as well as computationally-based evaluation techniques.

## References

[1] H. Christensen, T. Batzinger, K. Bekris, K. Bohringer, J. Bordogna, G. Bradski, O. Brock, J. Burnstein, T. Fuhlbrigge, R. Eastman *et al.*, "A roadmap for us robotics: from internet to robotics," *Computing Community Consortium*, 2009.

[2] R. Gockley, J. Forlizzi, and R. Simmons, "Interactions with a moody robot," in *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction*. ACM, 2006, pp. 186–193.

[3] D. Mazzei, N. Lazzeri, D. Hanson, and D. De Rossi, "Hefes: An hybrid engine for facial expressions synthesis to control human-like androids and avatars," in *Biomedical Robotics and Biomechatronics (BioRob), 2012 4th IEEE RAS & EMBS International Conference on*. IEEE, 2012, pp. 195–200.

[4] T. Hashimoto, S. Hitramatsu, T. Tsuji, and H. Kobayashi, "Development of the face robot saya for rich facial expressions," in *SICE-ICASE, 2006. International Joint Conference*. IEEE, 2006, pp. 5423–5428.

[5] F. Hegel, F. Eyssel, and B. Wrede, "The social robot flobi: Key concepts of industrial design," in *RO-MAN, 2010 IEEE*. IEEE, 2010, pp. 107–112.

[6] K. Berns and J. Hirth, "Control of facial expressions of the humanoid robot head roman," in *Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on*. IEEE, 2006, pp. 3119–3124.

[7] C. Benoit, J. Martin, C. Pelachaud, L. Schomaker, and B. Suhm, "Audio-visual and multimodal speech systems," *Handbook of Standards and Resources for Spoken Language Systems-Supplement*, vol. 500, 2000.

[8] A. Takeuchi and T. Naito, "Situated facial displays: towards social interaction," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM Press/Addison-Wesley Publishing Co., 1995, pp. 450–455.

[9] B. Le Goff, "Synthèse à partir du texte de visage 3 d parlant français," Ph.D. dissertation, 1997.

[10] J. Bassili, "Emotion recognition: the role of facial movement and the relative importance of upper and lower areas of the face." *Journal of Personality and Social Psychology*, vol. 37, no. 11, p. 2049, 1979.

[11] K. Gouta and M. Miyamoto, "Emotion recognition: facial components associated with various emotions]." *Shinrigaku kenkyu: The Japanese Journal of Psychology*, vol. 71, no. 3, p. 211, 2000.

[12] M. Ochs, R. Niewiadomski, C. Pelachaud, and D. Sadek, "Intelligent expressions of emotions," *Proceedings of the International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 707–714, 2005.

[13] D. Stork and M. Hennecke, *Speechreading by humans and machines: models, systems, and applications*. Springer, 1996, vol. 150.

[14] D. Berry, L. Butler, F. de Rosis, J. Laaksolathi, C. Pelachaud, and M. Steedman, "Final evaluation report," 2003.

[15] C. Pelachaud, "Some considerations about embodied agents," in *Proc. of the Workshop on Achieving Human-like Behavior in Iteractive Animated Agents, in the Fourth International Conference on Autonomous Agents*, 2000.

[16] L. Riek and P. Robinson, "Challenges and opportunities in building socially intelligent machines [social sciences]," *IEEE Signal Processing Magazine*, vol. 28, no. 3, pp. 146 –149, may 2011.

[17] C. Becker-Asano and H. Ishiguro, "Evaluating facial displays of emotion for the android robot geminoid f," in *Affective Computational Intelligence (WACI), 2011 IEEE Workshop on*. IEEE, 2011, pp. 1–8.

[18] N. Christoph, "Empirical evaluation methodology for embodied conversational agents," *From Brows to Trust*, pp. 67–90, 2005.

[19] L. Riek, P. Paul, and P. Robinson, "When my robot smiles at me: Enabling human-robot rapport via real-time head gesture mimicry," *Journal on Multimodal User Interfaces*, vol. 3, no. 1, pp. 99–108, 2010.

[20] B. Abboud and F. Davoine, "Bilinear factorisation for facial expression analysis and synthesis," in *Proceedings of the Vision, Image and Signal Processing, IEEE -*, 2005, pp. 327–333.

[21] Q. Zhang, Z. Liu, B. Guo, and H. Shum, "Geometry-driven photorealistic facial expression synthesis," in *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, 2003, pp. 177–186.

[22] M. Ochs, C. Pelachaud, and D. Sadek, "An empathic virtual dialog agent to improve human-machine interaction," in *Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems*, vol. 1, 2008.

[23] Y. Lee, D. Terzopoulos, and K. Waters, "Realistic modeling for facial animation," in *Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques*. ACM, 1995, pp. 55–62.

[24] B. Guenter, C. Grimm, D. Wood, H. Malvar, and F. Pighin, "Making faces," in *Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques*. ACM, 1998, pp. 55–66.

[25] H. Lee, J. Park, and M. Chung, “A linear affect–expression space model and control points for mascot-type facial robots,” *IEEE Transactions on Robotics*, vol. 23, no. 5, pp. 863–873, 2007.

[26] M. Ochs, R. Niewiadomski, and C. Pelachaud, “How a virtual agent should smile?” in *Intelligent Virtual Agents*. Springer, 2010, pp. 427–440.

[27] E. Bevacqua, D. Heylen, C. Pelachaud, M. Tellier *et al.*, “Facial feedback signals for ecas,” *in Proceedings of the AISB’07: Artificial and Ambient Intelligence*, pp. 328–334, 2007.

[28] J. Goetz, S. Kiesler, and A. Powers, “Matching robot appearance and behavior to tasks to improve human-robot cooperation,” in *Proceedings of the 12th IEEE International Workshop on Robot and Human Interactive Communication*, 2003, pp. 55 – 60.

[29] F. Rosis, C. Pelachaud, I. Poggi, V. Carofiglio, and B. Carolis, “From greta’s mind to her face: modelling the dynamics of affective states in a conversational embodied agent,” *International Journal of Human-Computer Studies*, vol. 59, no. 1, pp. 81–118, 2003.

[30] C. Kidd and C. Breazeal, “Effect of a robot on user perceptions,” in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2004, pp. 3559–3564.

[31] L. Riek, T.-C. Rabinowitch, P. Bremner, A. Pipe, M. Fraser, and P. Robinson, “Cooperative gestures: Effective signaling for humanoid robots,” in *5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2010, pp. 61 –68.

[32] L. Riek and P. Robinson, “Using robots to help people habituate to visible disabilities,” in *IEEE International Conference on Rehabilitation Robotics (ICORR)*, 2011, pp. 1 –8.

[33] H. Kobayashi, Y. Ichikawa, M. Senda, and T. Shiiba, “Realization of realistic and rich facial expressions by face robot,” in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2003, pp. 1123–1128.