How the agent's gender influence users' evaluation of a QA system

Andreea Niculescu, Dennis Hofs, Betsy van Dijk, Anton Nijholt
Human Media Interaction Group
University of Twente
Post BOX 217, 7500 AE, Enschede, The Netherlands
{niculescuai, D.H.W.Hofs, bvdijk, A.Nijholt}@ewi.utwente.nl

Abstract— **In this paper we present the results of a pilot study investigating the effects of agents' gender-ambiguous vs. gender-marked look on the perceived interaction quality of a multimodal question answering system. Eight test subjects interacted with three system agents, each having a feminine, masculine or gender-ambiguous look. The subjects were told each agent was representing a differently configured system. In fact, they were interacting with the same system. In the end, the subjects filled in an evaluation questionnaire and participated in an in-depth qualitative interview. The results showed that the user evaluation seemed to be influenced by the agent's gender look: the system represented by the feminine agent achieved on average the highest evaluation scores. On the other hand, the system represented by the gender-ambiguous agent was systematically lower rated. This outcome might be relevant for an appropriate agent look, especially since many designers tend to develop gender-ambiguous characters for interactive interfaces to match various users' preferences. However, additional empirical evidence is needed in the future to confirm our findings.**

Keywords: anthropomorphic agents, gender-ambigous vs. gender-marked look, user studies

I. INTRODUCTION

Physical characteristics, such as age, gender and ethnicity are important cues in human social perception, cognition and behavior [1]. They represent a kind of 'business card' that tells people how to approach a potential interaction partner. Research studies showed that humans prefer to engage in conversation with those whose physical appearance can be labeled consistently [2]. The reason is the human tendency to simplify the interlocutor's representation by framing them into pre-defined categories (e.g. old, male, Asian) [3]. This framing lightens the cognitive load and gives a secure feeling of dealing with predictable situations [4].

Among all salient visual cues that coin physical appearance, gender seems to be of fundamental importance, being part of the first visual information people exchange in daily communication. The explanation goes beyond the cognitive load lightening and relates to our evolutionary history, where gender related information assured the correct orientation toward a potential mating candidate. Since the decoding of such information has powerful impact on social interactions we believe that its lack would be perceived as unpleasant. In other words, we assume that humans would prefer to interact with those whose gender can be consistently labelled and would maintain this preference, even when interacting with artificial entities, such as avatars or conversational agents.

II. RELATED WORK

The virtual gender issue has become a highly interesting topic to the HCI community since many computer media systems started to use representative human avatars. Previous research has demonstrated that humans treat computers as if they were social actors, even though they do not exhibit anthropomorphic traits [5]. By adding a face and embodiment to an interface the social relationship between user and computer becomes even more explicit: clothing, facial expression, hairstyle, gender and age cues displayed by an agent bring the rich and complex world of human social interactions into the interface [6].

A number of researchers have studied the effects of 'virtual' gender on the way people perceive conversational agents and build relationships with them. Zimmerman et al. [7] concluded that people prefer agents displaying gender stereotypes conform to specific roles – female agents were preferred for tasks traditionally undertaken by women (librarian, matchmaker), male agents for tasks undertaken by men (fitness trainer). They also found that men prefer embodied agents more than women do and that female agents were preferred over male agents, by both male and female users.

Baylor et al. [8] investigated how the attitudes of female student undergraduates towards engineering were influenced by agents' age, gender, and "coolness". They found that, after interacting with a female agent, test subjects reported more positive stereotypes of engineers; after interacting with a male agent test subjects regarded engineering as being more useful.

Another study by Catrambone et al. [9] suggested that male and female test subjects might have different ways to personify agents. Their study showed that 54% of the female participants used a personal pronoun (he/she) to refer to an agent, while only 13% of the male participants did the same.

De Angeli and Brahnam [3] found the gender of the virtual embodiment impacts the incidence of sex talk: agents that clearly signalized their genders (female, male characters) were more prone to verbal abuse than those that did not do so (robot character).

However, with the exception of [3] there are no other studies known to the authors investigating human perceptions of gender-ambiguous characters. It is surprising, since often computer applications display avatars or agents whose physical appearance does not point to any particular gender. This look is intentionally created by designers with the purpose that both male and female users could relate to the character. Therefore, in this paper we propose a study investigating the impact of agents' gender-ambiguous vs. gender-marked look on the perceived interaction quality of a multimodal question answering (QA) system.

III. EXPERIMENT DESIGN

For our experiment we used the multimodal information system IMIX (Interactive Multimodal Information eXtraction) [10]. The system is an interactive QA engine for medical queries in Dutch (see figure 1): users can ask a medical question and get a response in the form of text and pictures, made-up by matching the query to document fragments from the system's database.

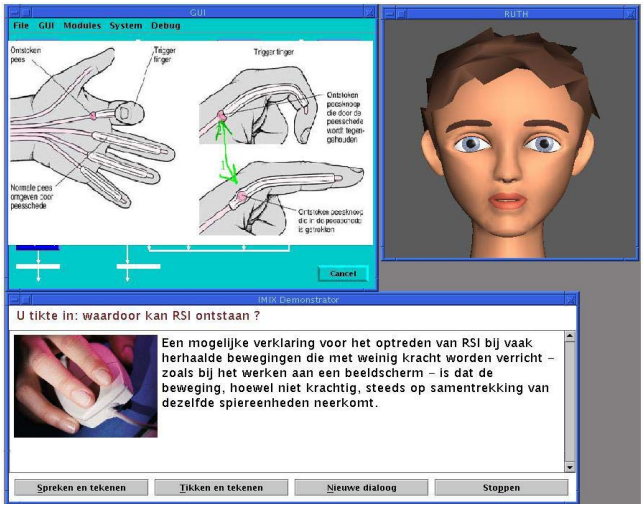

Figure 1. The IMIX system

Users can interact with IMIX using speech, text type or pen-like input (using the mouse). The answers are both presented in spoken language by a conversational agent and displayed on the screen. The dialogue with the system can also include greetings, clarification questions, feedback, and error handling strategies. Optionally, users can formulate follow-up questions in the form of text, speech or drawings.

The system's users are expected to be people with no professional knowledge of the medical domain. They would probably make use of IMIX occasionally. No special training is required to interact with the system.

The agent attached to the system is called Ruth[1]. Ruth could be classified as either a young male or a female with slight masculine traits (head on the far right, see figure 2). Additionally, in Dutch the name 'Ruth' designates a female person, thus potentially increasing users' confusion on Ruth's gender.

Based on Ruth we developed another two agents by adding simple but very explicit feminine, respectively masculine characteristics. The modifications were kept to a minimum in order to make the comparison between the agents sustainable - factors such as beauty, facial symmetry or hair color should not influence subjects' preferences for a particular agent.

The female agent, called Anna, is depicted wearing earrings, having narrow eyebrows, a lighter skin color and a feminine hairstyle; the masculine agent, called Bart has a slightly darker skin color, a beard and short hair. Pitch and frequencies of the synthesized TTS voices were adapted to fit the agents' gender; for Ruth, a female voice with grave pitches was chosen.

Figure 2. Feminine, masculine and gender -ambigous agent' heads

The visual perception of Ruth's gender ambiguity was confirmed by a preliminary study[2] performed with 48 male and female test participants, aged between 16 and 73 years and originating from 10 different countries across Europe, Latin America, Asia and Middle East. The participants were shown the images of seven different agents - among those was also the agent Ruth - and were asked to rate the agents' degrees of feminity and masculinity on a 5 point scale. No direct question addressed the gender ambiguity in order to avoid priming effects. We also counterbalanced the order in which the participants rated the agents to overcome potential biases. The results showed that Ruth's gender was perceived neither as masculine nor as feminine, placing the agent in the gender-ambiguous category. Detailed experimental settings can be found in [11].

IV. EXPERIMENT SET-UP

In the current experiment setting the subjects used only the speech modality to interact with the system, i.e. did not use the keyboard or the mouse. We chose the speech input option because it increases the naturalness of the interaction.

To ensure homogeneity between the trials we applied the Wizard-of-Oz technique and replaced the speech recognition module by a wizard. However, we introduced one simulated

[1] The head was developed at Rutgers University, New Jersey (USA): http://www.cs.rutgers.edu/~village/ruth/.

[2] No voice feedback was included in this study.

speech recognition error in each evaluation session to avoid the impression that the system was controlled by a human operator. A list with all questions was prepared in advance for the wizard. During the interaction the wizard could quickly copy-paste the questions in the QA interface minimizing the risk of delays or input mistakes that could have been caused by manual typing. The test subjects were informed that they were interacting with three differently configured systems, i.e. using different search algorithms. In fact, they are interacting with only one system controlled by the wizard. Each system was represented by a different gender-marked agent.

The test subjects received a set of three scenarios per evaluation session and accomplished a total of 9 (3 x 3) trials. The scenario sets (A, B, C) were constructed in a similar manner to provide equal conditions in terms of answer quality and time spent to complete the tasks (see table 1). Nevertheless, to overcome possible scenario weakness leading to a less positive system assessment we rotated the agents assigning them to a different scenario set each time.

Additionally, we randomized the order in which the participants interacted with the agents to exclude any potential biases that might arise from being exposed to one particular agent before the others.

TABLE I. SCENARIO SETS

Scenario No.	Set A	Set B	Set C
1.	1.What is the heart? 2.What represents the picture?	1.What is the lung? 2.What represents the picture?	1.What is the eye? 2.What represents the picture?
2.	1.What is hay fever? 2.What are the symptoms? 3.What causes hay fever? 4.How can hay fever be healed?	1. What is RSI? 2.What has RSI with stress to do? 3.What are the symptoms? 4.What represents the picture?	1.What is asthma? 2.What are the symptoms of asthma? 3.How can asthma be cured? 4.What represents the picture?
3.	1.What is the DNA? 2.What represents the picture? 3.What is a chromosome? 4.What represents the picture?	1.What is malaria? 2.What causes malaria? 3.What are the symptoms? 4.How can malaria be cured?	1.What is the sleeping sickness? 2.What are the symptoms? 3.How can the sleeping sickness be healed? 4.What represents the picture?

V. EVALUATION DESIGN

To determine whether test subjects perceive the interaction quality with the agents differently, we deployed two complementary evaluation methods: one quantitative short questionnaire to be filled in after each evaluation session and one in-depth qualitative interview conducted after the entire experiment.

A. Quantitative questionnaire

The purpose of the quantitative questionnaire was to give an idea of the preferences trend of the participants, i.e. no significant statistical results were meant to be achieved.

Our short survey was inspired by the SASSI questionnaire [12]. SASSI measures the usability of speech-based interfaces and addresses six dimensions: response accuracy, likeability, cognitive demand, annoyance, habitability and speed.

For each dimension – except for the speed – we chose one to three variables. We excluded the speed dimension because, according to our experiment settings, we were not expecting perceptible differences between the systems; nevertheless, we included this dimension in the qualitative interview.

Since the experiment was carried out within the limitation of a pilot study we used only 10 questions for the questionnaire and re-arranged the variables in two factor subscales: the first subscale measures interaction related features and contains five variables: mental load, interaction ease, response clarity, system flexibility and system efficiency; the second scale refers to interaction effects on users' mood; we called this subscale "user feeling" and it subsumes four variables: enjoyment, tenseness, degree of confidence and comfort. The survey ended with a question regarding the overall interaction quality (see table 2).

All variables were rated on a 20-point level scale to assure fine grained results.

TABLE II. STRUCTURE OF THE QUANTITATIVE QUESTIONNAIRE

Interaction features	User feelings	Interaction quality
1. mental load 2. interaction ease 3. response clarity 4. system flexibility 5. system efficiency	1. enjoyment 2. tenseness 3. degree of confidence 4. comfort	1.Overall interaction quality

B. Qualitative questionnaire

In the qualitative interview subjects were asked about their system preferences along several functional and non-functional interaction aspects involved in the experiment (see table 3). Under functional aspects we included questions about system transparency, response accuracy, response speed, response quality and feedback strategies. In the non-functional category we asked questions related to agent and interface aesthetics, voice quality, content formulation and trustworthiness. The interview ended with a question about the overall system preference.

TABLE III. STRUCTURE OF THE QUALITATIVE INTERVIEW

Functional aspects	Non-functional aspects	System preference
1. system transparency 2. response accuracy 3. response speed 4. response quality 5. feedback strategies	1. agent aesthetics 2. interface aesthetics 3. voice quality 4. content formulation 5. trustworthiness	1.Overall system preference

VI. RESULTS

The experiments and evaluation sessions lasted in total one hour. Additionally, the subjects were interviewed for another 25 minutes.

Eight test persons participated in the study: half of them were male, half were female. The small sample of participants was purposely chosen since we were intending to determine whether our study is worth pursuing.

Most of the subjects belong to the age group 20-30, except for two participants, whose ages were between 50-60 years. All participants - except for one - had a technical background and were knowledgeable about QA systems; half of the subjects had even used a QA system in the past.

A. Quantitative questionnaire

The reliability analysis performed on our subscale shows acceptable internal consistencies (Cronbach's α values are between 0.6-0.8, as shown in table 4).

TABLE IV. FACTOR SCALE AND MEAN AVERAGES

Factor category	Mean average		
	Anna	**Bart**	**Ruth**
Interaction features **No. items: 5** Cronbach's α: Anna =.816 Bart = .662 Ruth =.756	15.48	14.85	14.23
User feelings **No. items: 4** Cronbach's α: Anna =.761 Bart = .630 Ruth =.841	16.90	15.71	15.50
Interaction quality (overall)	14.75	14.00	13.75
TOTAL	**15.71**	**14.85**	**14.49**

The system represented by Ruth had the lowest mean average ratings for all factor categories. On the other hand, the system represented by Anna was systematically better rated.

To check the significance level of the mean differences we performed repeated ANOVA[3] measurements followed by a paired t-test with Bonferroni correction adjustments for multiple comparisons. The ANOVA measurements indicated statistically significant differences between the systems $F(73.29,1.0)=4.35$, $p<.05$ (the degrees of freedom were corrected using Huynh-Feldt estimates). The paired t-test revealed significant differences between the Anna system and the systems represented by Bart ($t= 2.38$, $p<.025$) and Ruth ($t= 2.46$, $p<.025$) on the factor "user feeling"; for the factors "interaction features" and "interaction quality" no statistically significant differences were found. Also, no significant differences were found between the evaluation scores given by female subjects and those given by male subjects.

[3] The data collected was normally distributed.

B. Qualitative interview

During the qualitative interview test subjects had the opportunity to talk openly about their experience with the IMIX system and its agents.

Regarding the system transparency, feedback strategies and general interface aesthetics no differences were found between the systems. Most of the test subjects (6 persons) considered the systems relatively transparent. The interaction style appeared to be intuitive and subjects knew right from the beginning how to handle the systems. The feedback was considered as to be sufficient by the majority (7 persons). Only one person complained about the reduced feedback visibility – some feedback statements were placed on the interface most top corner and could be easily overlooked. The interface, which was identical during all evaluation sessions – except for the agent's look - was considered as to be acceptable (6 persons) but relatively simple and containing only basic features.

However, regarding all other remaining aspects the test subjects did find differences between the systems. When asked about their system preference more than half of the test subjects (5 persons) chose the Anna system. Among the reasons for preferring Anna were mentioned the response accuracy (3 persons), response quality (4 persons) and, surprisingly, response speed (2 persons). Anna's answers appeared to be more nicely formulated, more informative and more relevant to subjects' queries (4 persons). The agent Anna was also considered as to have the pleasantest look (7 persons) and a much nicer voice (5 persons). Her look appears to be "more professional", like a "nurse" or a "teacher". Anna left the impression she was more knowledgeable, more trustworthy (7 persons) and more appropriate as a "medical expert" (7 persons), as compared with the other two agents.

Bart appeared less trustworthy because of his beard, while Ruth appeared to be too young and quite "dull". The Bart system was preferred by 2 test subjects in terms of answer quality (2 persons), response speed (2 persons), and agent look (1 person).

Only one single subject showed a general preference for the Ruth system but immediately added that Anna had a nicer face. Interestingly, even in situations where the content delivered by the Ruth system was, according to the participant's own statements, better – the participant still declared he would prefer the Anna system blaming the scenario setting for making Anna unable to give the desired answer (!).

VII. DISCUSSION

Despite the small number of test subjects our results are astonishing: even if test subjects interacted with the same system they felt significantly more comfortable, more confident and less tense with the Anna system, enjoying the interaction much more, as compared with the other systems. The Anna system also appeared to perform better than Bart or

Ruth; yet, this result could not prove statistical significance. During the qualitative interviews most of the test subjects confirmed their preference for the Anna system.

In general, test subjects seemed to prefer the gender-marked agents Anna and Bart over the gender-ambiguous agent Ruth, while Anna, got most of the preference "votes". Thus, our study shows encouraging results for our hypothesis concerning the human preference for interacting with consistently gender labeled entities.

On the other side we are aware of the fact that creating three different agents we created not only three different gender representation but also three different characters, i.e. the particular details of the three agents evaluated may subtly influence many dimensions of user attitude towards the agent beyond the gender. However, since gender effects cannot be studied purely, i.e. isolated from the face or voice of a character these influences are inherent. We believe by keeping the facial modification to a minimum to have reduced the unavoidable impact of these influences on our manipulations.

VIII. CONCLUSIONS

In this paper we presented the results of a pilot study performed with eight test subjects concerning the effects of agent's gender look on the perceived interaction quality of a multimodal QA system. In the future we plan to continue our research on this topic conducting other studies with a larger number of participants and an additional set of agents displaying similar gender-ambiguous vs. gender-marked characteristics in order to gain statistical evidence for our hypothesis.

REFERENCES

[1] M.A. Hogg, and G.M Vaughan, Social psychology. Harlow, Pearson Prentice Hall, 2005.

[2] C. Nass, K. Isbister, and E.J. Lee, "Truth is Beauty*:* Researching Embodied Conversational Agents", in Embodied Conversational Agents*,* J. Cassell, J. Sullivan, S. Prevost, E. Churchill, Eds. Cambridge, MA: MIT Press, pp. 374—402, 2000.

[3] A. De Angeli, and S. Brahnam, "Sex stereotypes and conversational agents". Proceedings of Gender and Interaction: real and virtual women in a male world, Venice, Italy, 2006.

[4] S. T. Fiske, and S. E. Taylor. Social Cognition, 2nd ed., New York: McGraw-Hill, 1991.

[5] B. Reeves, and C. Nass, The media equation: how people treat computers, television, and new media like real people and places. New York: Cambridge University Press/CSLI, 1997.

[6] J. Forlizzi, J. Zimmerman, V. Mancuso, and S. Kwak, "How interface agents affect interactions between humans and computers". Proceedings of Designing Pleasurable Products and Interfaces, ACM Press, pp. 209 – 221, 2007.

[7] J. Zimmermann, E. Ayoob, J. Forlizzi, and M. McQuaid, "Putting a face on embodied interface agents". Proceedings of Designing Pleasurable Products and Interfaces*,* Eindhoven Technical University Press, pp. 233--248, 2005.

[8] A. L. Baylor, R. B. Rosenberg-Kima and E. A. Plant, "Interface agents as social models: the impact of appearance on females' attitude toward engineering". Proceedings of CHI - Extended Abstracts, pp: 526—531, 2006.

[9] R. Catrambone, J. Stasko, J. Xiao, "Anthropomorphic agents as a UI paradigm: experimental findings and a framework for research", Technical Report GIT-GVU-02-10, 2002.

[10] M. Theune, B. van Schooten, R. op den Akker, W. Bosma, D. Hofs, A. Nijholt, E. Krahmer, C. van Hooijdonk, and E. Marsi, "Questions, pictures, answers: introducing pictures in question-answering systems". Proceedings of the 10th International Symposium on Social Communication, Santiago de Cuba, 2007.

[11] A. Niculescu, F. Sluis, and A. Nijhot, "Feminity, masculinity and androgyny: how humans perceive the gender of anthropomorphic agents". Proceedings of 13th International Conference on Human-Computer Interaction, Springer Verlag, Heidelberg, pp. 628-632, 2009.

[12] K. S. Hone, and R. Graham, "Towards a Tool for the Subjective Assessment of Speech System Interfaces (SASSI)", Natural Language Engineering, vol. 6 (3-4), pp. 287—303, 2000.