International Conference on Artificial Reality and Telexistence
Eurographics Symposium on Virtual Environments (2023)
J.-M. Normand, M. Sugimoto, and V. Sundstedt (Editors)

Insights From a Study on Subtle Mimicry in Human-Agent Interaction

Robin Ungruh[1], Susanne Schmidt[1], Nahal Norouzi[2], and Frank Steinicke[1]

[1]Universität Hamburg, Germany
[2]University of Central Florida, USA

Abstract

In social interactions, people tend to imitate the behavior of others and to perceive dialogues in which they are imitated to be more natural and smooth. This process of mimicry is not limited to non-verbal behavior, but also involves subtle adaptation of one's own speech style to the communication partner. Although being a natural phenomenon in human-human interaction, it is not yet common for virtual agents to simulate such behavior by adapting their speech style to that of the user.
This work presents a user study ($N = 48$) that explores the participants' perception of a virtual agent mimicking formal and informal speech. The majority of participants preferred agents with a matching speech style over those with a mismatching one. Other positive results of mimicry that were previously found in human-human interaction could not be replicated. To inform other researchers studying subtle agent behavior about possible factors that might dominate participants' perception of an agent, we present the results of a thorough content analysis of qualitative user feedback. From the salient themes, such as mismatched emotionality in language and speech, affordances of agents, and expectations of the agent's role in interactions, we derive recommendations for the design of future user studies of subtle (verbal and non-verbal) agent behavior.

CCS Concepts

• ***Human-centered computing*** → ***Empirical studies in HCI;*** *Natural language interfaces;* • ***Computing methodologies*** → ***Discourse, dialogue and pragmatics;***

1. Introduction

Natural communication between people has many facets, one important one being mimicry. Humans imitate aspects of gestures, facial expressions, and the speech of their conversation partners [CL13]. Verbal mimicry, the adaptation of (para)linguistic features of the speech to an interlocutor, has been shown to positively affect the perception of an interaction and to enhance pro-social behavior toward the mimicker [GO07,GJM09]. That is, we mimic the attributes of people we like, and like people more who mimic us.

Despite being a subconscious and natural behavior in human-human interaction, it is still uncommon for intelligent virtual agents (IVAs) to be able to adapt their speech style towards the users'. Implementing such behavior, however, could make agents even more valuable, especially in contexts that require the agent to interact with different users. For example, customer service agents could adjust to how formal or casual the user approaches them and respond accordingly. The potential for such subtle differences to influence users, such as impacting their perceptions of human likeness and likeability of agents, has not been thoroughly researched, and the few results to date are inconsistent (see Section 2.2).

Besides the subtle nature of mimicry effects, one factor contributing to the decidedly mixed results of previous studies may be that researchers generally strive to isolate very specific factors in order to understand the influences of those factors as variables, keeping everything else constant. While keeping things constant enables comparisons, those constants can have unintended influences and often are not studied thoroughly. In a review by Norouzi et al. [NKH*18], it was noted that very few studies on IVAs collected qualitative data that would allow for analyzing and understanding the perceptions and ratings of study participants.

This work is contributing to the research of human-agent interaction in two ways. First, we investigate effects of verbal mimicry by both embodied and voice-only agents. In contrast to previous studies, we particularly focus on the formality of speech, i.e. formal (respectful, structured) versus informal (colloquial, slang) language [Joo67] (for examples, see Table 1). It is known that the speech style of users, and particularly the level of formality, varies not only between individuals, but also for the same person when expressing an idea to different audiences, using different modalities, or accomplishing different tasks [HD99]. Therefore, mimicry of these variations has the potential to personalize a user's interaction with a virtual agent. In a user study with 48 participants, we quantitatively evaluate the hypothesis that an agent's mimicry

DOI: 10.2312/egve.20231313

of the user's formality has a positive effect on the sense of social presence, perceived anthropomorphism, rapport, and overall preference (see Sec. 3.6). Second, we explore some of the effects of choices that are commonly made in designing IVAs by performing a qualitative analysis on the collected user feedback (see Sec. 3.7). Based on these results, we discuss general considerations for future studies of subtle agent behavior (see Sec. 4).

2. Related Work

2.1. Verbal and Nonverbal Mimicry

Mimicry (also called mirroring and synchrony) is the automatic process of imitating the behavior and actions of conversation partners or observed individuals [CL13]. This process can occur fully unconsciously and is usually initiated a few seconds after observing the actions to be mirrored. Common examples of mimicry are the imitation of yawning, body postures, face touching, food consumption, and even micro-movements like finger touching (for an overview, see [CL13]). Mimicry is not limited to nonverbal behavior but can also be observed in written and spoken conversations. According to the *Communication Accommodation Theory* (CAT), humans converge their speech style in terms of linguistic features (e.g., accent and speech rate) and paralinguistic features (e.g., pauses between words, utterance length, and pitch) [GO07]. In addition, interlocutors tend to reuse single words, phrases, and structures of previous sentences in their speech [LK82, PK06]. As a consequence of mimicry, various positive effects on individuals were demonstrated, including increased liking, empathy, and affiliation in social interactions (for an overview of related studies see Chartrand et al. [CL13]).

To date, these effects have only been partially replicated for human-agent interactions. In a study by Bailenson et al. [BY05], mimicking users' head movements led to a significantly higher social presence (i.e., the feeling of being there with a "real" person [OBW18]) and a significantly more positive impression of the agent. Hale et al. observed a significantly stronger rapport (i.e. the feeling of connection and harmony with a conversation partner [HMG11]) towards an agent that mimicked the user's head and torso movements [HA16]. This effect, however, could not be replicated in a second study by the same authors [HA16]. Other hypothesized positive effects on trust, perceived similarity, smoothness of interaction, and self-other overlap were also not found. In another study on the effect of agents mimicking facial expressions, no effect on rapport was found [HVDSLG18]. Further results on verbal mimicry are discussed in the following section.

2.2. Formality-Based Speech Styles

Language style was defined by Enkvist [Enk16] as variations in the language that preserve the content to be conveyed. In 1967, the linguist Martin Joos [Joo67] introduced five different speech styles of the English language, characterized by their level of formality and associated with specific contexts (from most formal to most informal): frozen, formal, consultative, casual, and intimate style.

A different definition of conversation styles in face-to-face dialogues was introduced by Tannen [T*05], who observed distinctive conversation behavior between her New Yorker and non-New Yorker friends. These distinctions formed the basis for the formation of two different speech styles; the *High Involvement* is characterized by a reduced syntactic form, overlaps between speakers and short silences between sentences, while in the *High Consideration* style, speakers tend not to interrupt each other and speak in a structured manner. A study by Shamekhi et al. found that users prefer virtual agents that match their own speech style to either *High Consideration* or *High Involvement* [SCM*16]. However, no significant effects of matching speech style on agent trustworthiness or likability were detected. Other findings by Hoegen et al. indicate that users with a *High Consideration* style perceive agents with a matching conversation style as more trustworthy; a result that could not be shown for users with *High Involvement* style [HAMC19]. Neither study found evidence for an increased perceived quality of interaction with agents that use a matched speech style.

Joos' definitions of formality-based speech style levels [Joo67] show that the extremes, frozen and intimate style, may generally be inappropriate for virtual agents. Most current agents in the form of home assistants or service bots use consultative style, meaning they engage in bi-directional conversations and typically do not use slang or humorous references that require shared background knowledge with the user. They thus simulate the speech style that is most common in human-human interactions and is appropriate for everyday conversations, especially with strangers or in a work setting [Joo67]. Besides, formal style may also be suitable for users who utilize agents just for information retrieval and wish for straightforward answers. Additionally, casual style may be useful for users who aim for friendship-like talks and personify the agent strongly. Considering the differences between users' personalities and their expectations of an agent's behavior, mimicking formal and informal speech has the potential to create more human-like, individualized experiences.

3. User Study

To investigate whether the positive effects of mimicking formality in human-human conversations can be transferred to human-agent conversations and whether simulating this human trait increases the extent to which the agent is perceived as a human-like social being, we conducted a user study. We hypothesized that a cross-over interaction exists between formality of the user's speech and formality of the agent's speech:

- **(H1)** Matching speech style has a positive effect on *perceived social presence*.
- **(H2)** Matching speech style has a positive effect on *perceived agent anthropomorphism*.
- **(H3)** Matching speech style has a positive effect on *rapport with the agent*.
- **(H4)** Matching speech style is *preferred* over mismatching speech style.

The user study was conducted in two iterations. In the first iteration, study participants held a conversation with an embodied virtual agent, thereby incorporating not only the agent's appearance, but also the agent's nonverbal behavior in the form of facial expressions, gaze, and gestures as additional factors. Based on the participants' feedback and due to the fact that the majority of currently used agents, such as Apple's Siri and Amazon's Alexa, are

Table 1: *Three example utterances from the study, with either formal or informal speech style. The same examples were used in the post questionnaire, asking participants to rate which sentences would correspond more closely to their own speech style.*

Formal Speech Style	Informal Speech Style
Good day, I am pleased to meet you.	Hey, cool to meet you.
I am especially fond of jazz music.	I love jazz music.
I'd be glad to discuss such topics again someday.	Glad to talk again, anytime!

voice-only, we conducted a second study iteration without agent embodiment with a different group of participants.

For better structuring and comparability, we will present both iterations jointly and point out differences at the appropriate places. Overall, the study will be treated as a mixed design with the between-subject factor *agent representation*, and the two within-subject factors *user formality* and *agent formality*. In the following section, we introduce the study methodology that allowed us to vary the user formality within participants.

3.1. Method

Although IVAs are currently used predominantly as task-oriented, transactional assistants, research on agents with personality, an autobiographical background, or their own motivations suggests that they will increasingly take on the role of social entities (for an overview, see [NKH*18]). We wanted our scenario to reflect this view of agents, without being associated with a very formal or very informal speech style per se. Thus, we decided to place the user in an interview situation where the agent assumed the role of the interviewee. In four rounds of alternating interview topics, the user was presented with a script of questions to ask the agent (see Section 3.4). In each round, a different combination of user formality and agent formality was chosen for the questions in the script and the agent's answers, respectively (for examples, see Table 1). The scripts were based on pairs of informal / formal sentences taken from the *Music & Entertainment* data set of the *GYAFC* corpus [RT18]. To create coherent dialogues, the scripts were padded with transitional sentences in consultation with a native English speaker. The full dialogues can be found in the supplementary material. The interview was conducted remotely via a simulated Zoom meeting. An interview situation was chosen due to a number of expected favorable characteristics:

- Users can follow a script allowing us to change the users' speech style between conditions, thus comparing (mis)matching speech styles in a within-subject design.
- It is plausible that user and agent interact without knowing each other beforehand.
- The conversation flow can be fully controlled, which increases comparability within and between participants.
- It is plausible to talk to the agent within a Zoom conference (either with or without video).
- Conversations are theoretically long enough to observe mimicry (as opposed to today's transactional interactions with agents, e.g., concerning the time or outside temperature).
- The predefined script prevents the user from adapting the questions and thus isolates mimicry effects that are exclusively caused by the (mis)matching agent's responses.

The method of pre-scripting the dialogue between user and agent has already been used by Shamekhi et al. [SCM*16] for studying mimicry of *High Consideration/Involvement* speech styles.

3.2. Measures

Social Presence To measure social presence, we used a corresponding questionnaire according to Bailenson et al. [BBBL01]. Since we expected differences between the conditions to be subtle, we increased the granularity of the Likert scale to a range of 1 ("strongly disagree") to 7 ("strongly agree"). We further adjusted the first question, which originally asked whether users sensed the presence of another person in the room. Since our setting was a Zoom meeting, the question was transformed into: "I perceived that I was in the presence of another person in the meeting with me."

Anthropomorphism For the second measure, we used a subscale of the Godspeed questionnaire introduced by Bartneck et al. [BKCZ09]. It originally measures perceived anthropomorphism with five items, each contrasting 2 adjectives on a 5-point Likert scale. Since pretests showed that five levels are too constraining for the small effect sizes expected in our scenario, the scale was increased to seven levels. A value of 1 expressed artificial characteristics, while 7 represented a high degree of human likeness.

Rapport The third questionnaire is an adapted version of the Human-Agent Rapport Questionnaire (HARQ) [CAGP16]. It originally utilizes a 5-point Likert scale, with "strongly agree" indicating a high feeling of rapport. Again, we expanded this to a 7-point scale based on results of a pre-test. Furthermore, the item "I was paying attention to [the] way that [the] character responds to me and I was adapting my own behavior to it" (factor loading of 0.49) was removed from the questionnaire because, in our study design, participants were not supposed to freely adapt their wording.

3.3. Materials

We conducted a remote study, requiring participants to attend a meeting via the Zoom video conferencing software. The experimenter hosted the Zoom session and additionally joined the meeting on an external PC under the name "Louise". Via this account, the screen was shared to show a Unity application with the agent. At the bottom of the agent window, the user was presented with the current interview question.

Figure 1: *Embodied agent for the first study iteration [Eis18].*

For the first iteration of the study, we used an agent with a virtual 3D body representation (see Figure 1). The agent was implemented in the game engine Unity, version 2021.1.17f1, as described in the article by Schmidt et al. [SAS20]. It is based on a 3D scanned female head model by Eisko [Eis18], extended with stochastic gaze behavior (including focusing on points of interest, saccades, and blinks) [Kna21] and lip synchronization. The agent uses IBM Watson APIs to transcribe the user's speech input and synthesize the agent's speech responses. Dialogues are handled through the natural language interface *Watson Assistant*. The interviews are covered by a tree of 8 different response schemes representing the four interview topics with either the formal or informal speech style of the agent. If an input sentence can be classified into one of these schemes, the agent returns the matching response. Otherwise, the agent asks the participant to repeat the question.

In the second iteration of the study, we simulated a video conference in which all attendees turned off their cameras. Based on feedback from the first iteration, where participants reported a noticeable mispronunciation of multiple words, we changed not only the embodiment of the agent but also her voice. This was done in an attempt to minimize distracting factors. While the wording of the responses remained the same, they were synthesized using the Google Text-to-Speech (TTS) engine (voice *en-US-Wavenet-F*). The change led to a considerable improvement in pronunciation, which was supported by the participants' feedback after the second study iteration (see Sec. 3.7.2). Since two participants of the first study iteration reported that their input sentences were not understood several times, we decided to use a Wizard of Oz approach in the second iteration, i.e., the agent's response was manually triggered by the experimenter.

In summary, the two iterations of the study differed in terms of agent representation, with the first being embodied and using Watson TTS output, while the second was voice-only using Google TTS output. In terms of timing and sentence structure, the change from automatic to manually controlled dialogue flow did not yield any noticeable difference from the experience of the first-iteration participants who did not encounter any problems with the speech recognition.

3.4. Procedure

Before the study, the participants signed a consent form, informing them about the study and how their data would be processed. When joining the Zoom meeting, they were briefly told about the procedure of the study. It was explicitly mentioned that the user would interview four different agents without disclosing the focus of the study on the agents' speech style. Participants were then asked to fill out a demographic questionnaire. Afterward, the interaction with the IVA was demonstrated by the experimenter, who initiated a brief dialogue to introduce the interview. After the demonstration, the experimenter turned off their camera. In the main part of the study, each participant sequentially conducted four interviews with the agent, targeting music, art, cuisine, and movie directors. While the sequence of interview topics was fixed, the condition order with four combinations of user and agent formality was counterbalanced among participants. To reduce the impression that the agent was impersonating the same individual in all four conditions and was thus unnaturally changing the speech style, the agent greeted and said goodbye to the user in each condition and did not show any signs of recognizing the user. After each interview, which lasted approximately 1 to 1.5 minutes, the participants were asked to answer three questionnaires, as presented in Section 3.2, and an open question asking for further comments.

After all conditions were completed, a post questionnaire was presented asking for general information about the participants (e.g., their speech style, their experience with voice agents) and their overall impression of the agents (e.g., agent preference, whether differences in agent formality were noticed). To assess the speech style of participants in everyday conversations, they were asked which sentences from two given sets of (formal and informal) sentences would rather correspond to their own style. After the questionnaire was filled out, participants had the opportunity to give unstructured oral feedback on their experience, both on the agent in general and on their impressions of the individual conditions. The study lasted approximately 30 to 45 minutes.

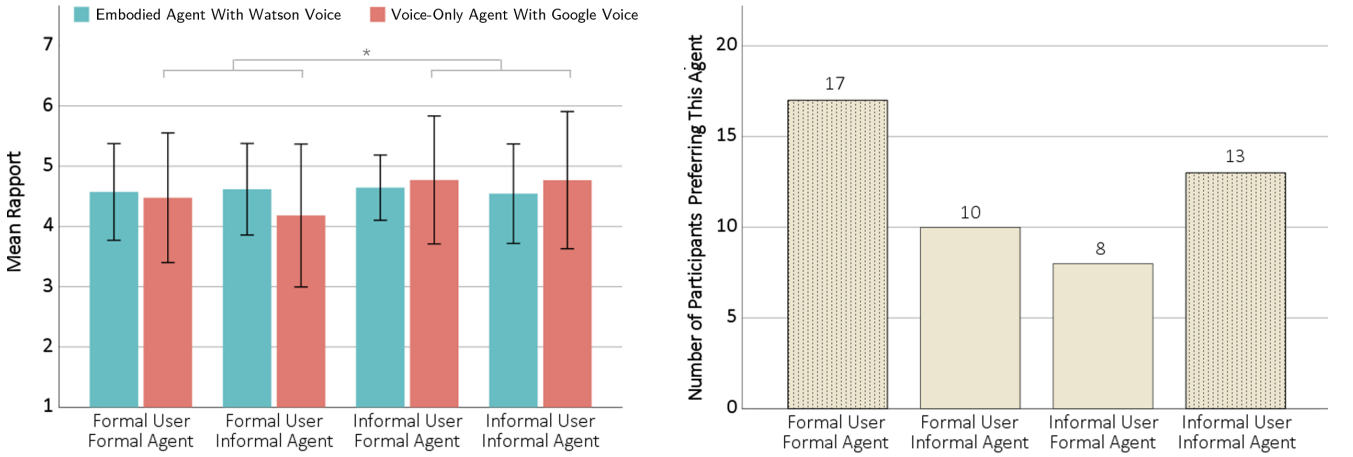
3.5. Participants

We invited 48 participants, 24 for the first iteration $i1$ with embodied agent (12 male and 12 female; ages between 21 and 31, $M = 25.0$) and 24 for the second iteration $i2$ with a voice-only agent (13 male and 11 female; ages between 22 and 56, $M = 27.8$). 43 participants were students or researchers in the field of Human-Computer Interaction (HCI) or Computer Science, and five students from other fields. HCI students were compensated with course credit. Four participants of $i2$ were native English speakers, while the remaining participants self-assessed their listening English skills in spoken interactions to be at a level of B1 ($N_{i1} = 2, N_{i2} = 1$), B2 ($N_{i1} = 4, N_{i1} = 3$), C1 ($N_{i1} = 10, N_{i2} = 6$) or C2 ($N_{i1} = 8, N_{i2} = 10$) as defined by the Common European Framework of Reference for Languages (CEFR) [Cou01]. Only three participants rated their general speech style as rather formal than informal.

3.6. Quantitative Analysis

3.6.1. Methods

The questionnaires were analyzed by calculating the means in every trial. Normal distribution of the scores' residuals can be assumed (assessed by Shapiro-Wilk tests and Q-Q plots). Therefore, a mixed



(a) (b)

Figure 2: *(a) Mean rapport values for both study iterations (vertical bars represent the standard deviation). The brackets indicate a significant simple main effect of user formality on rapport for the voice-only agent with Google voice. (b) Preferred condition with dotted bars representing matching speech styles.*

Table 2: *Means and standard deviations for each of the four conditions per study iteration.*

Agent Representation	**Embodied Agent With Watson Voice**								**Voice-Only Agent With Google Voice**							
User Formality / **Agent Formality**	**Formal** / **Formal**		**Formal** / **Informal**		**Informal** / **Formal**		**Informal** / **Informal**		**Formal** / **Formal**		**Formal** / **Informal**		**Informal** / **Formal**		**Informal** / **Informal**	
	M	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**	**M**	**SD**
Social Presence	3.68	1.02	3.77	1.20	3.68	1.15	3.59	1.09	3.64	1.32	3.43	1.35	3.73	1.37	3.86	1.40
Anthropomorphism	3.23	.86	3.13	1.08	3.20	.79	3.24	1.05	3.57	1.56	3.37	1.55	3.66	1.58	3.92	1.47
Rapport	4.57	.80	4.62	.76	4.64	.54	4.54	.83	4.48	1.08	4.18	1.19	4.77	1.06	4.77	1.14

ANOVA was performed for each measure, with *agent representation* as between-subject factor (with the levels *embodied* and *non-embodied*) and *user formality* as well as *agent formality* as within-subject factors (each with the levels *formal* and *informal*).

3.6.2. Results

Means and standard deviations for all variables are listed in Table 2. No significant main or interaction effects on the social presence and perceived anthropomorphism of the agent were found. For the rapport measure, the ANOVA revealed a significant interaction between agent representation and user formality ($F(1,46) = 5.967, p = .018, \eta_p^2 = .115$). A subsequent simple main effects analysis with Sidak-adjusted comparisons revealed that for the second study iteration (i.e., the voice-only agent), rapport scores were .441 points higher for informal than for formal user speech ($p = .001$). For the first study iteration (i.e., the embodied agent), no significant simple main effect on rapport was found ($p = .990$). The results are illustrated in Figure 2a.

In the final questionnaire, participants were asked about their preferred agent (see Fig. 2b). 30 participants selected a condition with matching speech styles (13 informal–informal and 17 formal–formal) while mismatching speech styles were favored 18 times (8 informal user–formal agent and 10 vice versa). One participant indicated that they liked each agent about equally, and another selected two conditions as being similarly favorable.

3.6.3. Discussion

Regarding our initial hypotheses, the collected data only supports **(H4)**. Ratings of social presence **(H1)**, agent anthropomorphism **(H2)** and rapport **(H3)** were not significantly higher for matching speech styles. Thus, results were obtained that are comparable to those of Shamekhi et al. [SCM*16], who also found a significant correlation between users' style and their preferred agent, but this was not reflected in ratings of the agent's likability or the desire to continue interacting with the agent (two items represented in a similar form in the HARQ).

Regarding user preference, while the majority of participants chose a condition in which the agent's responses had a speech style similar to the user's scripted questions, we observed a strong recency effect, with 20 participants choosing the last agent and only one participant choosing the first agent. One factor contributing to this result could be increasing familiarity with the agent, as ex-

plained in Section 3.7.2. Without a recency effect, the differences in preference for matching versus mismatching speech styles shown in Figure 2b can be expected to become even more pronounced, as the recency effect in the current study should have contributed equally to all four conditions due to the counterbalanced order of conditions among participants with 100% coverage of order effects.

In the second iteration of the study, we observed significantly higher rapport scores for an informal than for a formal user speech style. This correlation between the own speech style and the relationship with the agent seems surprising at first glance, but a look at the HARQ rapport questionnaire indicates a reason. Besides questions explicitly concerning the agent (e.g., *I think the character is likable*), there were also questions assessing the interaction as a whole (e.g., *I felt uncomfortable during the interaction*). Since 45 of the participants indicated that their usual speech style in everyday conversations is informal rather than formal, being constrained to read a formal script with predefined expressions may have had an impact on how comfortable they found the conversation, which in turn may have lowered the overall HARQ score.

One factor that may contribute to the missing support for **(H1)** – **(H3)** is that mimicry effects have been found to be subtle in previous human-agent interaction studies [HAMC19, HA16, VHPM13] and thus may require a larger sample size to be detectable. Furthermore, greater differences between conditions were found in the second study iteration, which contained fewer potentially distracting factors than the embodied agent in study iteration 1, both due to the missing agent body and the higher quality of the synthesized voice. Indications of whether participants were actually distracted from the conversation by other sources may be found in qualitative feedback from study participants in the open-ended questions and their oral feedback.

3.7. Content Analysis

To gain insight into factors that may have influenced the perception of and interaction with the agent, we conducted a content analysis of the qualitative comments collected throughout the user study.

3.7.1. Methods

We collected all written responses to the following open-ended questions:

- *Is there anything else that you noticed or that you would like to mention?* (asked after each of the four conditions)
- *Can you explain why [you liked the previously chosen agent best]?* (asked once in the post questionnaire)
- *Have you noticed differences in the agent's degree of formality while talking?* (asked once in the post questionnaire)
- *Any other observations or comments that you would like to share?* (asked once in the post questionnaire)

This resulted in overall 191 non-empty responses, 153 of which contained qualitative feedback (rather than, e.g., just "yes"). Each response was transferred to a sticky note in the web-based whiteboard platform *Miro* [Mir21]. We then followed a content analysis approach, with inductive (i.e., observation-based) coding of the data [Arm17]. Comments were clustered by common themes. If a comment targeted different aspects of the experience, it was split into multiple notes. Comments that were ambiguous or did not address a specific aspect of the interaction (e.g., *Hope to have such agent in my life*) were discarded ($N = 19$). After an initial iteration, similar clusters were merged and the resulting themes were spatially arranged to reflect semantic proximity. Based on the clustering, several core themes were derived (see Table 3). In the final step, each note was reviewed with respect to the final categorization and reassigned in individual cases. All steps were performed by the first two authors and required agreement between both.

3.7.2. Results

The resulting categorization with an example utterance and the determined frequencies per category is shown in Table 3. Our content analysis revealed multiple factors that influenced participants' perceptions of the agent and therefore presumably affected ratings of social presence, anthropomorphism, and rapport. The first type of derived themes concerns dimensions of the agent (representation), ranging from high-level qualities that are unique to embodied agents, such as appearance and behavior, to basic qualities that also apply to chatbots, such as the expressed content.

Behavior Particularly in the first trials, participants often focused on the agent's eye movements, which were occasionally perceived as unnatural, even though they followed an advanced model based on human motion data [Kna21]. Interestingly, the lack of direct eye contact was mentioned, although this is also common in real video conferences using a standard technical setup [TWL20, HXX21]. Negative comments were also made about the agent not smiling or showing other facial expressions, while the synchronized lip movements and subtle breathing animation were positively emphasized.

Appearance While only mentioned 3 times in the written comments, the majority of participants in the first study iteration reported in their oral feedback that the agent's appearance contributed to their overall ratings. It was noted that initially much attention was paid to the visualization of the character and that the differences in speech were not noticed until the later trials.

Speech The agent's voice was one of the most salient factors participants commented on, with individually varying perceptions such as *artificial*, *robotic*, but also *natural* and *human-like*. Although the voice was consistent across all four conditions, participants noted differences in perceived naturalness as well as other features such as speaking rate. Switching from Watson to Google TTS seems to have improved perceived quality, as no more pronunciation errors were mentioned in the second study iteration.

Linguistic Style Variations in speech style, which are the actual focus of our study, were also noticed by most participants (36 according to the final questionnaire). In the open-ended feedback, participants indicated three different contexts in which matching of speech style had a positive effect, that is, when the agent mimics (i) the user's script, (ii) the user's actual speech style, and (iii) what the user perceives as normal in general everyday conversations. Individual phrases were specifically highlighted if participants felt that those were not part of their own usual vocabulary.

Content Content-wise, participants expressed favor for the

Table 3: *Categorization of the participants' utterances in open-ended questions related to the perception of the agent. S_1 refers to the first study iteration with an embodied agent, and S_2 to the second iteration with the voice-only agent.*

Category	Example	Count S_1	Count S_2
Behavior	"She doesn't look me straight in the eye."	14	–
Appearance	"I liked her face."	3	–
Speech	"The voice still sounds artificial, but already good enough to be enjoyable to listen to."	11	14
Linguistic Style	"She sometimes uses very strong adjectives that seem unnatural."	14	11
Content	"The agent and I share the same interests."	3	9
(Mis)matching Aspects	"The robo-voice combined with informal speech patterns make up for a very funny image."	11	3
Perceived Personality	"Seemed approachable, friendly and polite."	18	11
Interaction with User	"Felt that the conversation was the most fluid one."	9	20
Affordances of Agents	"Interesting that the agent was talking about what food they like [...] I can't imagine a computer eating anything."	2	7
Study Design	"Would be cool if you had some freedom in the questions asked."	4	7

general conversation if they could relate to the topic being discussed, and for the agent if they had the same opinion on the topic.

(Mis)Matching Aspects A recurring theme in the open-ended feedback was inconsistencies between the aforementioned agent dimensions, which can be observed in similar forms within state-of-the-art applications featuring virtual agents [COC21, YUY21]. Specifically, it was mentioned that enthusiastic responses by the agent were often accompanied by unemotional facial expressions and intonation, and that the informal speech style seemed inappropriate for the agent's appearance and voice.

Besides these comments directly referring to specific qualities of the agent, we found multiple high-level themes that cannot solely be assigned to a single dimension of the agent, but are formed through their combination, subjective interpretation and/or relation to general expectations of the user towards agents.

Perceived Personality Based on the overall impression of the previously described aspects, a variety of character traits were attributed to the agent, including *polite*, *lively*, *relaxed*, *warm*, and *humorous*, but also *distant*, *egotistical*, and *arrogant*. Some participants even empathized with the agent to assess whether she enjoyed the conversation. The fact that the agent expressed her own opinion was also rated positive in some cases and strange in others.

Interaction with User The general flow of conversation was a frequently mentioned evaluation factor. Specifically named aspects were the (unbalanced) conversation shared between participant and agent (since the agent's answers were usually more elaborate than the user's questions) as well as whether participants felt that the agent listened to their opinion and asked follow-up questions. Regarding the evolution of the human-agent interaction over time, six participants explicitly indicated that they became accustomed to the situation and therefore perceived the agent as less of a computer but more natural or even more likable.

Affordances of Agents Several participants commented that the agent's capabilities and status differed from their expectations. This related firstly to the fact that the agent commented on taste, which, unlike vision and hearing, is not a sense that users would attribute to a computer. Moreover, based on previous experience, agents have been portrayed as a tool that provides information and to which users issue commands rather than as a human-like entity that expresses opinions and preferences (e.g., Amazon Alexa, Google Assistant, and Apple Siri).

Study Design Finally, there were some general suggestions about the design of the study. A few participants expressed the desire to have open conversations with the agent instead of following a script, but from the experimenter's point of view, this would have introduced additional confounding factors.

3.7.3. Discussion

The amount and variety of themes revealed by our content analysis supports our hypothesis from Section 3.6.3 that participants' evaluation of the agent was influenced by many factors other than the formality of the speech style. Specifically, only 13% of the comments were related to the agent's speech style. Implications from these findings, as well as approaches to counteract them in future studies, are presented in the following section.

4. Insights into Studying Subtle Agent Behavior

Based on the preceding analysis of user feedback, we discuss limitations of our design choices as well as some general options that can be considered in the conception of future human-agent interaction studies, especially when subtle changes in the agent's (verbal or non-verbal) behavior are involved.

Interview scenario Our study design was adopted from previous work on the mimicry of speech styles, with regard to both the pre-scripted conversation and the personal content of the conversation (e.g., having the agent refer to her sense of taste) [SCM*16]. However, the significant main effect of user formality on HARQ

scores detected in our analysis may indicate that study participants felt uncomfortable reading a script that did not match their usual speech style. Related research has also found significant differences between spontaneous and read speech, both in terms of acoustic and linguistic features [NIF08]. An alternative approach to elicit varying speech styles in an unscripted manner might be to place study participants in different scenarios. Formal language, for example, could be naturally provoked by pretending that the interaction is being observed by a high authority (e.g., a professor). In this case, however, it would be necessary to investigate whether expectations concerning the agent's speech style actually originate from the user's own speech style or from the situation itself.

Another observation regarding the study design was that although study participants conducted an interview with the agent, several criticized the agent for having personal experiences and an opinion of their own. Unequal conversational shares and the fact that the agent did not ask any questions back to the user were even perceived as arrogant or egoistic. This suggests that the interview scenario with the agent in the position of the interviewee was not perceived as natural by the users, possibly because it changes the usual power dynamic between users and agents, with the latter normally appearing as personal assistants. Also, users might still tend to perceive agents as task-oriented, functional entities.

Natural Appearance, Speech, and (Emotional) Behavior At the time of implementing the presented study, both text-to-speech engines and facial animation systems targeting non-professionals still faced difficulties in expressing natural behavior with a rich range of emotions. Particularly in the first iteration of the study with an embodied agent and IBM Watson-generated speech, several participants perceived the agent's appearance and behavior as artificial or incongruent. Recent developments in these areas, such as Google's *SoundStorm* [BSV*23] for natural audio generation or *Ziva Face Trainer*[†] for enhanced facial rigs, suggest that such artificiality and incongruity can be reduced in future studies, which would be beneficial in the study of subtle agent behavior.

Special attention should be paid to the IVA's gaze, which was one of the most prevalent themes in our content analysis. In our embodied agent, eye movements were determined stochastically, rarely leading to situations where the agent would look away when the user began to speak, which was perceived as irritating or even impolite. A future gaze model should integrate typical behaviors known from human-human communication, such as maintaining eye contact when the conversational partner is speaking.

With regard to conversational flow, where interruptions can be very obvious to users as indicated by the qualitative feedback, switching to a Wizard of Oz setup brought an improvement in the case of our study. With the recent introduction of large generative pre-trained transformer (GPT) models [BMR*20], a transition back to an intelligent dialogue system should be considered for future studies, as this would allow more freedom in the conversation with the agent, as requested by some of our study participants. However, further consideration of inter-condition confounding factors due to non-deterministic response behavior of agents is necessary.

[†] https://zivadynamics.com/ziva-face-trainer

Interaction Duration Due to the limited duration of each encounter between user and agent (1 to 1.5 minutes in our study), the agent appears as a stranger to the participants, possibly influencing the user's attitude toward it. Apart from a long-term study, which is often not feasible, a familiarization period could reduce the impact of this limitation in future studies. Our content analysis suggests that another positive effect of such a phase is that users become accustomed to the agent's voice and/or appearance, making the agent perceived as less artificial over time and directing the user's attention away from aspects initially perceived as unnatural.

Context Dependency Like many related IVA projects (e.g., [BY05, HA16, VHPM13]), we selected a specific agent appearance and voice for our study. However, our qualitative results suggest a strong influence of the interaction context conveyed by the entirety of the scenario, from clothing style to additional objects in the scene to the assumed age of the agent. Since it is challenging to create a fully generic scenario (e.g., our agent's wool sweater was perceived as formal by some participants), it might be helpful to have a future use case scenario in mind when designing the virtual study environment. Another option, propagated mainly in psychology research, is called stimulus sampling [MO14, WW99]. Here, small aspects are varied between the participants and treated as random factors in the analysis. Stimulus sampling in terms of agent appearance could also support the impression that users are interacting with four different agents rather than one agent that unnaturally exhibits a different speech style in each study condition.

5. Conclusions

In this paper, we presented a user study ($N = 48$) investigating the mimicry of formal and informal speech by (dis)embodied virtual agents. While the majority of study participants preferred conversation with an agent that matched the participant's speech style, this result was not reflected in ratings for perceived anthropomorphism or rapport. A content analysis of user responses suggests that variations in formality were perceived by most participants, but contributed little to the overall ratings due to several other influencing factors. Based on these results, we discussed multiple practical approaches aimed at reducing such influences in future IVA studies.

Beyond adapting the study design to follow these approaches, it would be interesting to extend mimicry research to cover additional questions. Our study had a focus on the influence of mimicry through the agent, but due to the concept's bidirectionality, the user would presumably also adapt to the agent's speech style over time. Corresponding observations have already been made when some participants converted contracted forms in their interview questions (e.g., *it's*) into the long form (e.g., *it is*) when speaking to a formal agent. Moreover, as indicated by user feedback, the perception of IVAs is highly dependent on their visual appearance and voice. Further research on this topic may be warranted to develop a generic agent that can accommodate different speech styles. Finally, machine learning language models such as GPT-4 already support human-like dialogue with a chatbot, with the option to automatically convert formal to informal language and vice versa. By coupling this with a user speech formality classifier, a more flexible approach can be developed for both future user studies and real-world applications.

References

[Arm17] ARMBORST A.: Thematic proximity in content analysis. *Sage Open 7*, 2 (2017), 1–11. 6

[BBBL01] BAILENSON J. N., BLASCOVICH J., BEALL A. C., LOOMIS J. M.: Equilibrium theory revisited: Mutual gaze and personal space in virtual environments. *Presence: Teleoperators & Virtual Environments 10*, 6 (2001), 583–598. 3

[BKCZ09] BARTNECK C., KULIĆ D., CROFT E., ZOGHBI S.: Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International journal of social robotics 1*, 1 (2009), 71–81. 3

[BMR*20] BROWN T. B., MANN B., RYDER N., SUBBIAH M., KAPLAN J., DHARIWAL P., NEELAKANTAN A., SHYAM P., SASTRY G., ASKELL A., ET AL.: Language models are few-shot learners. *arXiv preprint arXiv:2005.14165* (2020). 8

[BSV*23] BORSOS Z., SHARIFI M., VINCENT D., KHARITONOV E., ZEGHIDOUR N., TAGLIASACCHI M.: Soundstorm: Efficient parallel audio generation. *arXiv preprint arXiv:2305.09636* (2023). 8

[BY05] BAILENSON J. N., YEE N.: Digital chameleons: Automatic assimilation of nonverbal gestures in immersive virtual environments. *Psychological science 16*, 10 (2005), 814–819. 2, 8

[CAGP16] CEREKOVIC A., ARAN O., GATICA-PEREZ D.: Rapport with virtual agents: What do human social cues and personality explain? *IEEE Transactions on Affective Computing 8*, 3 (2016), 382–395. 3

[CL13] CHARTRAND T. L., LAKIN J. L.: The antecedents and consequences of human behavioral mimicry. *Annual review of psychology 64* (2013), 285–308. 1, 2

[COC21] CLARK L., OFEMILE A., COWAN B. R.: Exploring verbal uncanny valley effects with vague language in computer speech. In *Voice Attractiveness*. Springer, 2021, pp. 317–330. 7

[Cou01] COUNCIL OF EUROPE. COUNCIL FOR CULTURAL COOPERATION. EDUCATION COMMITTEE. MODERN LANGUAGES DIVISION: *Common European Framework of Reference for Languages: learning, teaching, assessment*. Cambridge University Press, 2001. 4

[Eis18] EISKO: Animatable digital double of louise by eisko©, 2018. URL: www.eisko.com. 4

[Enk16] ENKVIST N. E.: *Linguistic stylistics*. De Gruyter Mouton, 2016. 2

[GJM09] GUEGUEN N., JACOB C., MARTIN A.: Mimicry in social interaction: Its effect on human judgment and behavior. *European Journal of Social Sciences 8*, 2 (2009), 253–259. 1

[GO07] GILES H., OGAY T.: Communication accommodation theory. In *Explaining communication: Contemporary theories and exemplars*, Whaley B. B., Samter W., (Eds.). Lawrence Erlbaum Associates Publishers, 2007, pp. 293–310. 1, 2

[HA16] HALE J., ANTONIA F. D. C.: Testing the relationship between mimicry, trust and rapport in virtual reality conversations. *Scientific reports 6*, 1 (2016), 1–11. 2, 6, 8

[HAMC19] HOEGEN R., ANEJA D., MCDUFF D., CZERWINSKI M.: An end-to-end conversational style matching agent. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents* (2019), pp. 111–118. 2, 6

[HD99] HEYLIGHEN F., DEWAELE J.-M.: Formality of language: definition, measurement and behavioral determinants. *Interner Bericht, Center "Leo Apostel", Vrije Universiteit Brüssel 4* (1999). 1

[HMG11] HUANG L., MORENCY L.-P., GRATCH J.: Virtual rapport 2.0. In *Intelligent Virtual Agents: 10th International Conference, IVA 2011, Reykjavik, Iceland, September 15-17, 2011. Proceedings 11* (2011), Springer, pp. 68–79. 2

[HVDSLG18] HOEGEN R., VAN DER SCHALK J., LUCAS G., GRATCH J.: The impact of agent facial mimicry on social behavior in a prisoner's dilemma. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents* (2018), pp. 275–280. 2

[HXX21] HE M., XIONG B., XIA K.: Are you looking at me? eye gazing in web video conferences. *methods 27* (2021), 28. 6

[Joo67] JOOS M.: *The five clocks*, vol. 58. New York: Harcourt, Brace & World, 1967. 1, 2

[Kna21] KNABE T.: Realistic eye movements. https://assetstore.unity.com/packages/tools/animation/realistic-eye-movements-29168, 2021. Accessed: 2022-05-13. 4, 6

[LK82] LEVELT W. J., KELTER S.: Surface form and memory in question answering. *Cognitive psychology 14*, 1 (1982), 78–106. 2

[Mir21] MIRO: The visual collaboration platform for every team: Miro. https://miro.com/, 2021. Accessed: 2022-05-13. 6

[MO14] MONIN B., OPPENHEIMER D. M.: The limits of direct replications and the virtues of stimulus sampling. 8

[NIF08] NAKAMURA M., IWANO K., FURUI S.: Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance. *Computer Speech & Language 22*, 2 (2008), 171–184. 8

[NKH*18] NOROUZI N., KIM K., HOCHREITER J., LEE M., DAHER S., BRUDER G., WELCH G.: A systematic survey of 15 years of user studies published in the intelligent virtual agents conference. In *Proceedings of the 18th international conference on intelligent virtual agents* (2018), pp. 17–22. 1, 3

[OBW18] OH C. S., BAILENSON J. N., WELCH G. F.: A systematic review of social presence: Definition, antecedents, and implications. *Frontiers in Robotics and AI 5* (2018), 409295. 2

[PK06] PARRILL F., KIMBARA I.: Seeing and hearing double: The influence of mimicry in speech and gesture on observers. *Journal of Nonverbal Behavior 30*, 4 (2006), 157–166. 2

[RT18] RAO S., TETREAULT J.: Dear sir or madam, may i introduce the gyafc dataset: Corpus, benchmarks and metrics for formality style transfer. *arXiv preprint arXiv:1803.06535* (2018). 3

[SAS20] SCHMIDT S., ARIZA O., STEINICKE F.: Intelligent blended agents: Reality–virtuality interaction with artificially intelligent embodied virtual humans. *Multimodal Technologies and Interaction 4*, 4 (2020), 85. 4

[SCM*16] SHAMEKHI A., CZERWINSKI M., MARK G., NOVOTNY M., BENNETT G. A.: An exploratory study toward the preferred conversational style for compatible virtual agents. In *International Conference on Intelligent Virtual Agents* (2016), Springer, pp. 40–50. 2, 3, 5, 7

[T*05] TANNEN D., ET AL.: *Conversational style: Analyzing talk among friends*. Oxford University Press, 2005. 2

[TWL20] TAUSIF M. T., WEAVER R., LEE S. W.: Towards enabling eye contact and perspective control in video conference. In *Adjunct Publication of the 33rd Annual ACM Symposium on User Interface Software and Technology* (2020), pp. 96–98. 6

[VHPM13] VERBERNE F. M., HAM J., PONNADA A., MIDDEN C. J.: Trusting digital chameleons: The effect of mimicry by a virtual social agent on user trust. In *International Conference on Persuasive Technology* (2013), Springer, pp. 234–245. 6, 8

[WW99] WELLS G. L., WINDSCHITL P. D.: Stimulus sampling and social psychological experimentation. *Personality and Social Psychology Bulletin 25*, 9 (1999), 1115–1125. 8

[YUY21] YORGANCIGIL E., URGEN B. A., YILDIRIM F.: Uncanny valley effect is amplified with multimodal stimuli and varies across ages. 7